UNIVERSITÉ DE MONTRÉAL

RELIABILITY AND QUALITY OF SERVICE IN OPPORTUNISTIC SPECTRUM ACCESS

ARASH AZARFAR
DÉPARTEMENT DE GÉNIE ÉLECTRIQUE
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

THÈSE PRÉSENTÉE EN VUE DE L'OBTENTION
DU DIPLÔME DE PHILOSOPHIÆ DOCTOR
(GÉNIE ÉLECTRIQUE)
JUILLET 2014

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Cette thèse intitulée :

RELIABILITY AND QUALITY OF SERVICE IN OPPORTUNISTIC SPECTRUM ACCESS

présentée par : AZARFAR Arash
en vue de l'obtention du diplôme de : Philosophiæ Doctor
a été dûment acceptée par le jury d'examen constitué de :

M. NERGUIZIAN Chahé, Ph.D., président
M. FRIGON Jean-François, Ph.D., membre et directeur de recherche
Mme SANSÒ Brunilde, Ph.D., membre et codirectrice de recherche
M. GENDREAU Michel, Ph.D., membre
M. LE Long Bao, Ph.D., membre

To my lovely parents...

ACKNOWLEDGEMENT

With no doubt, this research work could not be completed without the substantial support, instructive comments and precious advice of my research supervisors Professors *Jean-François Frigon* and *Brunilde Sansò*. Their positive outlook in my research inspired me and gave me confidence. I would like therefore to express my sincere gratitude to them.

I would also like to thank my doctoral defense jury, Professors Chahé Nerguizian, Michel Gendreau and Long Le, for dedicating time to my thesis and for their valuable comments. I also thank Professor André Girard for his instructive comments in my qualification exam.

I am deeply grateful to my friends and also all colleagues at Professor Frigon's research group and Professor Sansò's LORLAB. Discussions with them have always been instructive and motivating. In particular, many thanks to Dr. Mohsen Eftekhari Hessari. I would also like to thank Dr. Mohammad Torabi, Saeed Ghazanfari Rad, Mostafa Mehrtash, Marnie Vondou, Esmaeil Naderi, Ehsan Bolouki, Mohammad Ali Torabi, Diego E. Perea, Farshad Sarabchi, Filippo Malandra, Vida Vakilian, Hadhami Dbira, Eddy Dailleux, Silvia Boiardi, Luca Gianoli and Dr. Hakim Mellah.

I would also like to acknowledge all staffs of the Department of Electrical Engineering and Poly-Grames Research Center, particularly Prof. Ke Wu, Dr. David Dousset, Ms. Nathalie Lévesque, Ginette Desparois, Elena Pavlov, Louise Clément and Ghyslaine Ethier-Carrier for their assistance with all the administrative work, and Mr. Jean Bouchard and Jean-Sébastien Décarie for their technical support related to computer systems. I also appreciate kind support of the staffs of GERAD, particularly Ms. Marie Perreault, Carole Dufour and Francine Benoit.

My special thanks also go to Professor Danijela Cabric and all the members of Cores at UCLA, specially Chun-Hao Liu, for their kind support and hospitality during my stay in UCLA. I am also grateful to all colleagues at Sciswave specially Mr. Samer Abielmona and Michael Davies.

I would like to acknowledge the financial support for this project from the Natural Sciences and Engineering Research Council of Canada (NSERC) and *Fonds de recherche du Québec Nature et technologies* (FQRNT).

I would like to take this opportunity to also thank my former professors at Sharif University of Technology, particularly my supervisor Dr. Amir Hossein Jahangir, for their support and encouragement.

I owe my deepest gratitude to my lovely family, my kind parents to whom I dedicate this thesis for their unconditional love and support, and to my lovely brothers and sister. Words cannot express how much I love them and how grateful I am for their support.

RÉSUMÉ

Les réseaux radio-cognitif constituent une des meilleures options technologiques pour les réseaux sans-fil futurs. Afin d'étudier comment la fiabilité devrait être redéfinie dans ces réseaux, nous étudions d'abord les sources les plus fréquentes de panne dans les réseaux sans-fil et fournissons une procédure systématique de classement des pannes. Il est ensuite expliqué comment les radios cognitives peuvent profiter de leur propre capacité à mettre en œuvre des mécanismes efficaces de prévention et de récupération contre les pannes et ainsi assurer des communications sans-fil fiables et de qualité de service constante. En considérant des normes arrivantes sur la base de l'OSA, ce qui distingue un réseau radio-cognitif de ses prédécesseurs est des changements fréquents de canal ainsi que de nouvelles exigences telles la détection de disponibilité et la décision d'utilisation du spectre. Nous nous concentrons sur cet aspect et modélisons la remise du spectre comme une panne. Par conséquent, améliorer la fiabilité est équivalent à augmenter le temps moyen entre pannes, à rendre plus efficace le processus de récupération et à réduire le temps moyen de réparation.

Nous étudions donc d'abord l'impact du temps de récupération sur la performance du réseau radio-cognitif. En classifiant les pannes en dures et souples, il est examiné comment la disponibilité, le temps moyen entre pannes et le temps moyen jusqu'à la réparation sont touchés par le procès de récupération. Nous observons que le temps dépensé pour la récupération empêche le réseau d'atteindre le maximum de disponibilité. Par conséquent, pour obtenir un temps plus élevé entre pannes et un temps de réparation plus court, une option disponible est d'augmenter le nombre de canaux pouvant être utilisés par le réseau radio-cognitif, de sorte que, avec une haute probabilité, un utilisateur qui a raté le canal puisse trouver bientôt un nouveau canal. De l'autre côté, un mécanisme de récupération efficace est nécessaire pour mieux profiter de ce grand nombre de canaux ; l'amélioration de la récupération est donc indispensable.

Pour étudier l'impact de la récupération sur les couches plus hautes (e.g., la couche liaison et réseau), l'approche de l'analyse de file d'attente est choisie. Compte tenu des périodes de récupération comme une interruption de service, un modèle général de file d'attente de M/G/1 avec des interruptions est proposé. Différents paramètres de fiabilité et de qualité de service peuvent être trouvés à partir de ce modèle de file d'attente pour étudier comment la spécification des canaux, tels la distribution des périodes de disponibilité et d'indisponibilité, et la spécification de l'algorithme de récupération, tels la durée de récupération, affectent les paramètres de performance comme la perte de paquets, de retard et de gigue, et aussi le temps entre pannes.

Pour soutenir la différenciation des classes de trafic, nous proposons une approche de file d'attente avec priorité. Nous proposons une extension des résultats du modèle de file d'attente générale et présentons quatre différentes disciplines de file d'attente de priorité, allant d'un régime préemptif absolu à un régime complètement non préemptif. Les nouvelles disciplines augmentent la flexibilité et la résolution de décision et permettent au noeud CR de contrôler l'interaction des différentes classes de trafic avec plus de précision. Les modèles de files d'attente sont analysés et sont résolus, donc nous pouvons discuter comment la fiabilité et la qualité de service d'une classe spécifique de trafic sont affectées non seulement par les paramètres du canal, mais aussi par les caractéristiques des autres classes de trafic.

La file d'attente M/G/1 avec des interruptions est une fondation pour l'analyse des performances et une réponse à la nécessité d'avoir des relations analytiques de forme fermée. Nous étendons ensuite le modèle de file d'attente pour des scénarios plus réalistes, d'abord avec les canaux hétérogènes (le taux de service étant hétérogène dans la file d'attente en fonction du canal sélectionné) et ensuite avec plusieurs utilisateurs et un modèle d'accès au médium de transmission aléatoire. Dans la première partie, l'occupation de la file d'attente est modélisée avec une chaîne de Markov multi-lignes où chaque ligne représente l'un des taux de services possibles. En plus de résoudre numériquement la chaîne de Markov, deux approximations analytiques sont fournies. Bien qu'un modèle de chaîne de Markov nécessite de supposer une distribution exponentielle pour la disponibilité des canaux, nous analysons plus loin et discutons le modèle de file d'attente de l'OSA avec la distribution générale du temps de service et des périodes de disponibilité. Les résultats d'analyse et de simulations indiquent que les paramètres comme le taux d'occupation moyen de file d'attente sont similaires pour les différentes distributions de temps de service et des périodes de disponibilité et que les modèles de Markov simplifiés (avec les distributions sans mémoire) peuvent être utilisés pour prédire avec précision les performances du trafic des réseaux OSA hétérogènes.

Un scénario multi-utilisateur en temps discret est également étudié pour caractériser l'impact d'un protocole d'accès au médium de transmission aléatoire avec canal de contrôle sur la performance des réseaux radio-cognitif. Il nous permet d'étudier l'interaction des algorithmes de récupération et le protocole d'accès au support. Nous observons que, avec un protocole d'accès au médium de transmission de type Aloha, la performance est plus élevée quand une politique de recouvrement d'attente est utilisé. Autrement dit, au lieu d'une politique de récupération de commutation et de changement de canal, c'est à dire, la libération du canal en cas d'apparition des utilisateurs principaux (titulaire de licence), une politique de recouvrement d'attente est utilisée. Cette politique implique que l'utilisateur CR attend que l'utilisateur principal quitte le canal. Dans un protocole d'accès au médium de type Aloha, le goulot d'étranglement est l'accès au canal de contrôle, et une politique de commutation nécessite des accès plus fréquents.

Bien que l'effort principal de ce travail de recherche soit d'analyser l'impact des processus de récupération sur la performance, à la fin, un mécanisme glouton de commutation du spectre basé sur l'apprentissage est suggéré pour améliorer le temps de récupération dépensé pour trouver un nouveau canal. Au début de chaque intervalle de temps et dépendamment sur l'état du canal en cours, le système calcule le nombre optimal de canaux devant être mesurés dans cette période de récupération, et ce nombre est dynamiquement mis à jour après chaque itération de récupération (chaque nouveau canal qui est mesuré). Les caractéristiques intrinsèques de l'apprentissage de la radio cognitive sont utilisées afin de créer une liste optimale de canaux à être évalués sur la base de l'information historique. Les résultats des simulations montrent que le mécanisme proposé améliore la performance de récupération en fournissant un temps de récupération plus court ou un canal restauré avec une qualité supérieure.

ABSTRACT

Cognitive-radio based wireless networks are a technology of choice for incoming wireless networks. To investigate how reliability should be redefined for these networks, we study the most common sources of failure in wireless networks and provide a systematic failure classification procedure. It is then explained how cognitive radios can use their inherent capabilities to implement efficient prevention and recovery mechanisms to combat failures and thereby provide more reliable communications and consistent quality of service in wireless networks. Considering incoming OSA-based standards, what distinguishes a cognitive radio network from its predecessors is the frequent spectrum handovers along with new requirements such as spectrum sensing and spectrum usage decision. We thus focus on this aspect and model the spectrum handover as a failure, so improving the reliability is equivalent to increasing the mean time to failure, improving the recovery process and shortening the mean time to repair.

We first study the impact of the recovery time on the performance of the cognitive radio network. By classifying the failures into hard and soft, it is investigated how the availability, mean time to failure and mean time to repair are affected by the recovery time. It is observed that the time spent for recovery prevents the network from reaching the maximum availability. Therefore, to achieve a high mean time to hard failure and low mean time to repair, an available option is to increase the number of channels, so that with a high probability, a user who missed the channel can soon find a new channel. On the other side, an efficient recovery scheme is required to better take advantage of a large number of channels. Recovery improvement is thus indispensable.

To study the impact of recovery on higher communication layers, a queueing approach is chosen. Considering the recovery periods as a service interruption, a general M/G/1 queueing model with interruption is proposed. Different reliability and quality of service parameters can be found from this queueing model to investigate how channel parameters, such as availability and unavailability periods, and the recovery algorithm specifications, such as the recovery duration, affect packet loss, delay and jitter, and also the MTTF and MTTR for hard and soft failures.

To support traffic differentiation, we suggest a priority queueing approach. We extend the results of the general queueing model and discuss four different priority queueing disciplines ranging from a pure preemptive scheme to a pure non-preemptive scheme. New disciplines increase the flexibility and decision resolution and enable the CR node to more accurately control the interaction of different classes of traffic. The models are solved, so it can be

analyzed how the reliability and quality of service parameters, such as delay and jitter, for a specific class of traffic are affected not only by the channel parameters, but also by the characteristics of other traffic classes.

The M/G/1 queueing model with interruptions is a foundation for performance analysis and an answer to the need of having closed-form analytical relations. We then extend the queueing model to more realistic scenarios, first with heterogeneous channels (heterogeneous service rate for different channels) and second with multiple users and a random medium access model. In the first part, the queue occupancy is modeled as a multi-row Markov chain where each row represents one of the possible service rates. In addition to numerically solving the Markov chain, two analytical approximations are provided. Although a Markov chain model necessitates assuming exponentially distributed channel availability periods, we further analyze and discuss the OSA queuing model for general distribution of service time and availability periods. The analytical and simulation results indicate that for usual system parameters, the queue average occupancy is similar for different distributions of service time and availability periods and that the exact memoryless Markov models can be used to accurately predict the heterogeneous OSA system traffic performance.

A multi-user scenario with discrete-time distributions is also investigated to have more insights on the impact of a baseline random medium access protocol and design of control channel on the performance of the cognitive radio networks. It enables us to study the interaction of recovery algorithms and the medium access protocol. It is observed that with an Aloha-type medium access, the performance can be better when instead of a switching recovery policy (i.e., vacating the channel in case of appearance of primary users), a waiting and buffering recovery policy is employed (i.e., the CR user waits for the primary user to vacate the channel) because in an Aloha-type medium access, the bottleneck is access to control channel and a switching policy necessitates more frequent control channel accesses.

While the main effort of this research work is to analyze the impact of spectrum handover recovery process, we also propose, a greedy and history-aware spectrum handover scheme to improve the time spent for spectrum handover (i.e., the recovery time). At the beginning of each timeslot and based on the state of the current channel, the scheme computes the optimal number of channels to be sensed in this restoration period and this number is dynamically updated after each channel sensing result. Intrinsic features of learning and history-awareness of CRs are used to create an optimal list of channels to be sensed based on the channels' background and historical information. Simulation results show that the history-aware sensing order improves the restoration mechanism by providing a shorter restoration time or a restored channel with a higher quality.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS AND ABBREVIATIONS

AMC	Adaptive Modulation and Coding
AP	Access Point
APS	Automatic Protection Switching
BS	Base Station
BSS	Basic Service Set
CDF	Cumulative Distribution Function
CPE	Customer-Premises Equipment
CR	Cognitive Radio
CRN	Cognitive Radio Network
eNB	(LTE) E-UTRAN Node B (Evolved Node B)
GDB	Geolocation Database
GDD	Geolocation Database Dependent
GPS	Geographical Positioning System
HP	High Priority
LP	Low Priority
MAC	Medium Access Control
MHCRN	Multi-hop Cognitive Radio Network
MPLS	Multiprotocol Label Switching
MTTF	Mean Time To Failure
MTTR	Mean Time To Recovery (Repair)
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
PDF	Probability Density Function
TTF	Time To Failure
TTR	Time To Recovery (Repair)
TVWS	TV White Space
RLSS	Registered Location Secure Server
WLAN	Wireless Local Area Network

CHAPTER 1

INTRODUCTION

1.1 Definitions and basic concepts

Wireless services have recently enjoyed tremendous success because users increasingly appreciate the ability to access or share information anywhere and anytime. In return for these conveniences, users have accepted that wireless links are unreliable with inconsistent quality of service (QoS) in which problems (dropped or hung connections, variable data rates, long delays, etc.) are frequent occurrences. Although users currently consider these problems as inherent characteristics of wireless networks, as wireless services will become more pervasive and replace applications currently provided only in wireline networks, the question of reliability and quality of service will eventually become more critical.

To provide a higher level of reliability and quality, more adaptable and intelligent wireless nodes are required, which are able to analyze the traffic requirements and adjust themselves accordingly. Cognitive radio (CR) was an answer to such a need Mitola III (2000). A cognitive radio node is an intelligent radio empowered with features such as spectrum-awareness, location-awareness, learning and history-awareness, adaptability and reconfigurability, and reasoning and decision-making capabilities. A CR can observe and learn from the environment and adapt its communication parameters based on this knowledge Mitola III (2000). This technology has been proposed as an extension to software defined radio (SDR) Dillinger *et al.* (2003); Jondral (2005); Ulversoy (2010) where this software architecture provides reconfigurability for CR to adjust its communication parameters based on its knowledge and observations. A collection of CR nodes that organize themselves using these cognitive features is called a cognitive radio network (CRN). CRN is a general concept and includes any wireless network from a point-to-point link between two stationary cognitive nodes to a complex cellular or mobile ad-hoc network.

The last fact regarding the improvement of quality of service in wireless networks is that while limited bandwidth and interference are two main impairment factors, the spectrum is also a limited and expensive resource in wireless communications. From 2001 to 2004, reports by the Federal Communications Commission (FCC) stated that on one side, due to extension of the wireless applications and mobile telephony networks, some parts of the spectrum are over-utilized and new requests for spectrum have dramatically increased (spectrum scarcity), while on the other side, several assigned frequency bands are under-utilized and partially used (spectrum usage inefficiency). To solve these problems, in the framework of dynamic spectrum access Zhao et Sadler (2007), the idea of using spectrum holes or spectrum opportunities to deploy new networks was proposed. This concept is often referred as opportunistic spectrum access (OSA).

As illustrated in Figure 1.1, a spectrum hole is a portion of the spectrum which is licensed to a primary network or user, but is vacant in a specific time and geographic area Haykin (2005) and can be temporarily used by a secondary network or user. This secondary manner of communication implies that a secondary user is able to use a vacant channel when the primary (licensed) user is not present. In wideband technologies, the 'vacant' term means that the interference level experienced by primary users will be acceptable (lower than a required threshold) if a secondary network starts operating in this channel. Cognitive radio (CR), based on its spectrum-awareness capabilities, was selected as the best candidate for implementing this concept. A secondary CRN consists of wireless nodes that are spectrum-aware. Using their spectrum-sensing capabilities, they detect chunks of unused spectrum licensed to primary users (e.g., television channels with no broadcaster in the geographic area) and deploy a secondary CRN in the available spectrum.

Meanwhile, it is expected that OSA implemented by cognitive radios, due to its comprehensive and powerful features, will be widely deployed in next generation wireless networks Jondral (2007); Mitola (2009), such as in local area and wide area networks Sherman *et al.* (2008). As discussed earlier, one of the main prerequisites for such a wide deployment is the support of higher *reliability* and *quality of service*. Altogether they motivate this research work on investigation of *quality of service and reliability in cognitive radio networks* with a quick glance at differentiation. The main two research directions of this thesis are thus evaluating and analyzing a cognitive radio network reliability or quality of service, and then providing solutions to improve the performance. Figure 1.2 represents the trend which results in the necessity of research in the area of reliability and quality of service in CR wireless networks.

The necessity to have a glance at differentiation is the fact that the deployment of wireless services in diverse applications and networks implies that wireless links may carry different kinds of traffic with variable requirements and importance ; thus, those networks should be able to adjust themselves to those dissimilarities of traffic and applications. Wireless networks should thus be able to provide differentiated levels of quality of service for different kinds of traffic and services, which raises the notion of *differentiated quality of service* in wireless networks. Let us explain it with two examples :

– The first example targets the replacement of DSL Internet access by wireless links. Nowadays, Internet access, besides its traditional roles, is providing many interactive

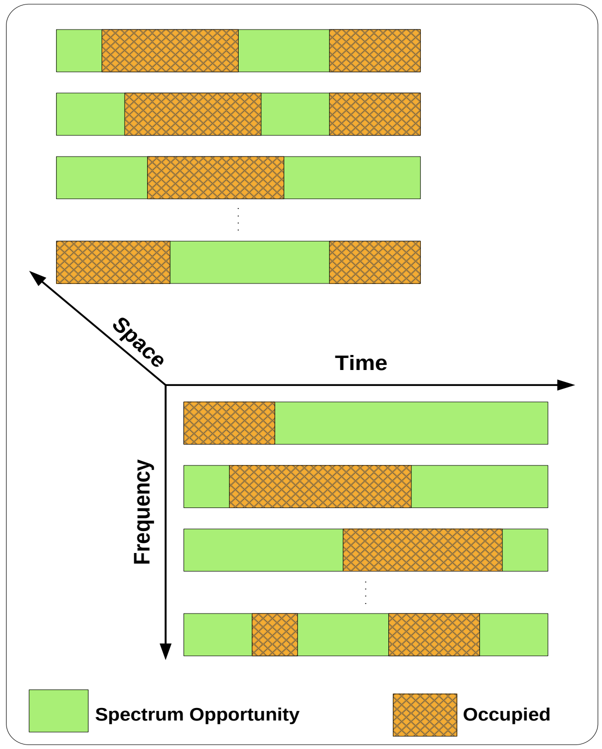

Figure 1.1 Spectrum Opportunity (SOP).

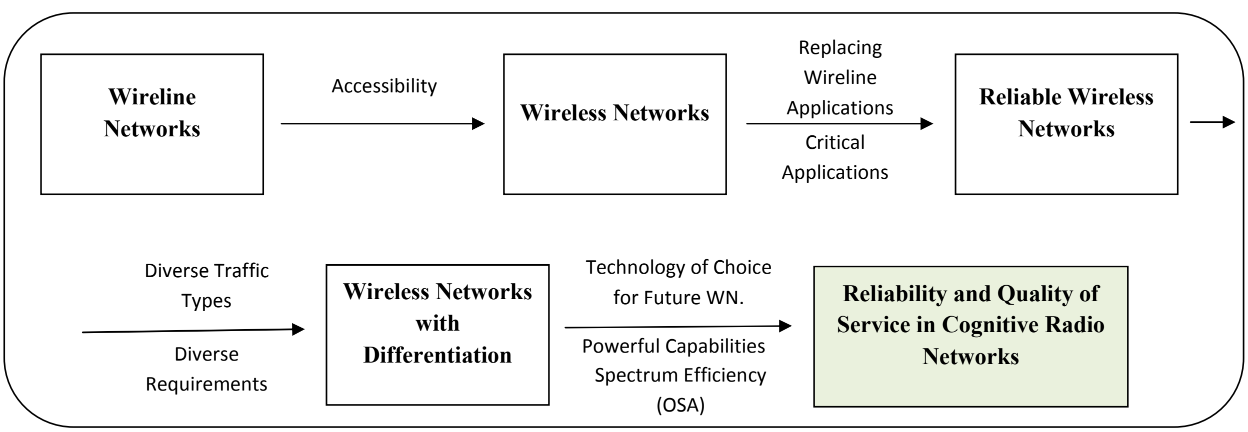

Figure 1.2 The necessity of research in the area of reliability and quality of service analysis in cognitive radio wireless networks.

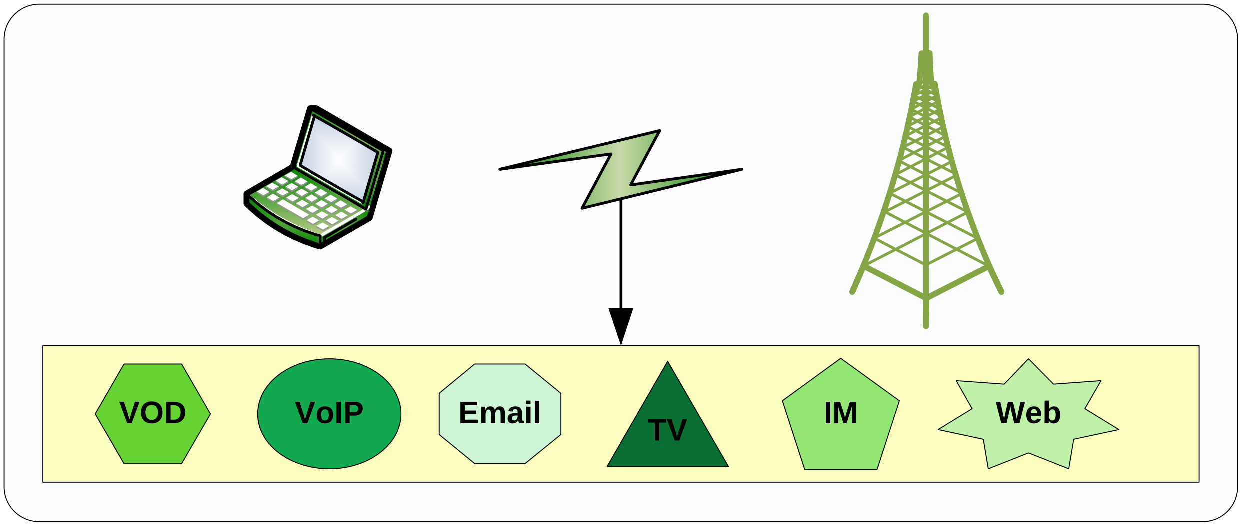

Figure 1.3 Nowadays, a wireless link carries different types of traffic.

services and is replacing air-broadcasting radio and TV channels and even gradually cable TVs. Therefore, on one side, wireless internet links will carry a bunch of traffic types with completely different requirements (please see Figure 1.3) and on the other side, users will not accept frequent disruptions during their favorite show in a live Internet TV program. Therefore, providing different levels of quality of service for these diverse application types over wireless access links is necessary.

– The second example is about the use of wireless networks in an emergency network with mission critical applications where the importance of reliability is more evident. An emergency network should have a very short setup time which implies a wired network could not be an appropriate candidate for such an application. The emergency network provides the possibility of communication between the disaster area and the disaster management center when a catastrophe might destroy all communication infrastructures. Over this network, besides the periodic data such as the report of temperature, humidity and the level of chemical pollution of the disaster area, voice messages and video reports will be transmitted. Certainly, while some lost packets of periodic data might be tolerated, an interruption and disconnection during the voice or video communication is not acceptable. The low service capacity of an emergency network besides the environmental hazards implies that this network should be equipped with some mechanisms to provide higher level of quality of service for more critical applications.

Nowadays, as discussed, the most important objective when employing cognitive radios is in the framework of opportunistic spectrum access to solve the problem of spectrum scarcity and spectrum usage inefficiency. With this introduction, two questions may arise. The first question is that while the CR technology seems to be a technology for lower communication layers, probably, the Internet Protocol (IP) will be still present in the network layer. Thus, why do not we trust existing quality of service solutions and differentiation mechanisms in the IP layer and leave this task to the IP Differentiated Services (DS) or other proposed methods for reliability improvement in communication networks ?

This question has been discussed in Sansò *et al.* (2006). The authors state that even though the IETF Diffserv can provide some level of reliability, as it is not designed for failure management and reliability purposes, its reaction to failures is slow and not efficient and consequently is not powerful enough to provide alone (differentiated) reliability. For this end, some collaboration between the IP and lower communication layers (in Sansò *et al.* (2006), they have mostly focused on the physical layer in WDM networks) is necessary, which results in cross-layer mechanisms to provide reliability in communication networks.

The other question that may arise after the above discussion is that when some other models have been proposed for the sake of quality of service improvement and differentiation, why not apply the same models to CRNs. In other words, what is the necessity of a new research on this topic in CRNs ?

The answer to this question can be divided into two parts. In the first part, we can mention the basic differences between a wired and wireless network which will be discussed in more details in Chapter 3. The problem is that, because of their nature, wireless networks are inherently more unreliable and error-prone. In addition to the basic hazards, including natural disasters, power failures and hardware failures, that threaten all communication networks, several other factors, such as the random nature of the communication channel and the presence of interferers, affect the reliability in wireless networks. The study of wireless networks reliability is therefore more challenging compared to the wired networks.

For the next part of the answer, we can highlight the distinction of CRNs compared to traditional wireless networks. This implies that even if there are some proposed models in the literature for the notion of reliability and quality of service improvement in wireless networks, they will not be powerful enough to handle this notion in CRNs. The first reason is that CR solves the problem of spectrum scarcity and usage inefficiency by opportunistically using the spectrum licensed to a primary network. OSA makes the resources unpredictable and time-varying which intensifies, at the first level, the unreliability of wireless communication in comparison to traditional wireless networks. As soon as the licensee returns to a channel, the CR network should vacate the channel. As one of the main differences which distinguishes a cognitive radio network from its predecessors is frequent spectrum handovers, we focus on this aspect and model the spectrum handover as a failure, so improving the reliability is equivalent to improving the mean time to failure, improving the recovery process and shortening the mean time to repair.

The second reason is that a CRN is also empowered with several intelligent and cognitive features which can be exploited to open new doors for the notion of reliability in wireless networks Azarfar *et al.* (2012c). As discussed in Azarfar *et al.* (2012c), any cognitive feature may help a CRN to prevent the occurrence of potential failures or to decrease their severity and consequence after occurrence, in a more efficient way. Therefore, ignoring these new capabilities and depending only on existing proposals is not an intelligent choice. This research thus deals with these cognitive features to investigate how they can be employed in a CRN to provide reliability and to improve the quality of service. These cross-layer proposals consider both unpredictability of the resources on one side and the notable cognitive capabilities of CR on the other side in order to guarantee the specified level of quality of service for different priority classes from an operational point of view.

1.2 Model Summary

As the thesis is being written in an article-based format, it is not easy to provide a detailed global model for the whole thesis. However in this section, we briefly explain the parts of the model which are common among different chapters. As illustrated in Figure 1.4, it is assumed that a cognitive radio node operates over its designated channel, known as current operating channel, for a duration of time. The operating time can be the time until the channel becomes unavailable, a fixed intended time (e.g., a timeslot) or the time until transmitting a packet. The last case is where, for instance, after transmitting each packet, the user should compete or decide for a new channel. Each operating period is followed by an interruption (recovery) period. The recovery time can be the time spent to decide and select a new channel, or the waiting time until the same channel becomes available, etc. The recovery time may include competition time, sensing and switching time, negotiation time, etc. Looking at the operation of the CR user from a reliability point of view, these recovery periods are the periods in which the link is being repaired therefore they represent *time to repair/recovery* in a reliability model. Considering Eq. (1.1) for the link availability, availability can be increased by decreasing the mean time to repair (MTTR), the duration of recovery periods, or increasing the mean time to failure (MTTF).

$$\text{Availability} = \frac{MTTF}{MTTF + MTTR}. \tag{1.1}$$

Once a channel is selected, the duration of the operating periods, equivalent to *time to failure*, is in most scenarios dictated by the environment, e.g., appearance of licensees, and is generally out of the control of the CR network. We thus mostly focus on the recovery time and the channel selection to investigate its impact on the performance and the way that it can be improved.

As explained in Figure 1.4, when the recovery time represents the time spent on sensing a list of channels and finding a new appropriate channel, the recovery time is a function of the order of the channels in the list, known as *channel sensing order*. Given that channels have a different probability of being available, different sensing time and different service capacity, questions such as which channels, how many channels and in which order they should be sensed, and which channel(s) should be selected (when the recovery is finished) determine the quality of the recovery process. Efforts on improving the recovery are concentrated on improving the spectrum decision algorithm.

While what we discussed above is general and for a single class of traffic, when applicable, multiple classes of traffic are considered and reliability analysis or improvement is investigated for multiple classes of traffic, in a differentiated manner.

1.3 Research objectives

The general objective of this research work is to investigate how reliability and quality of service can be improved in CRNs. Considering the reliability modeling in cognitive radio networks (to be discussed in Chapter 3), increasing the reliability is equivalent to decreasing the probability of failure occurrence (increasing the mean time to failure) and decreasing the recovery time. We focus on the latter approach, thus the objective is mostly to optimize the recovery process ; i.e., finding the best channel in the shortest time. To be able to evaluate a proposed recovery approach, mathematical tools and equations which relate the performance and the recovery process parameters are required. Since we found a lack of such tools in the literature, a new objective arose to analyze the performance (reliability or quality of service) in presence of the recovery periods. The specific objectives of the research work presented in this thesis can thus be listed as :

- Redefinition of reliability (failure, recovery, etc) in cognitive radio networks :
  - How should the reliability and quality of service be defined ?
  - How can it be improved ?
- Performance analysis of a cognitive radio network in presence of interruptions (recovery periods) :
  - Reliability analysis ;
  - Quality of service analysis ;
- Optimization of the recovery process for multiple classes of traffic :

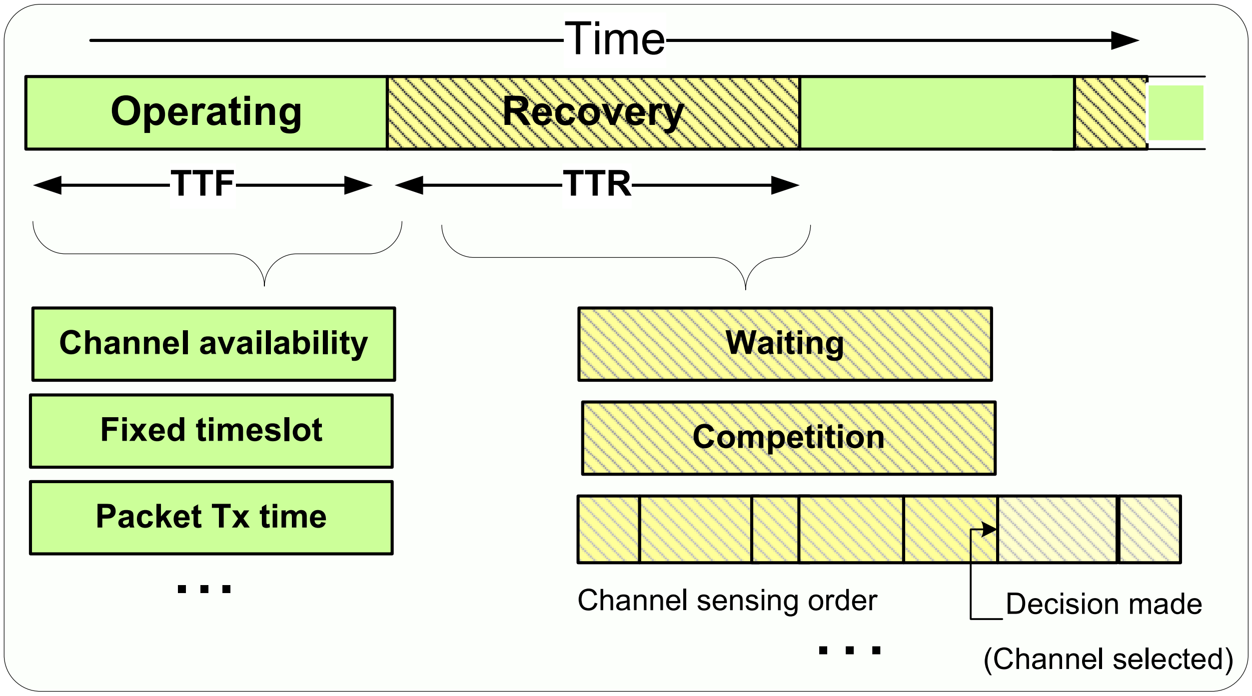

Figure 1.4 Operation and interruption (recovery) periods of a cognitive radio link.

– Applying priority ;
– Optimal sensing orders.

Figure 1.5 illustrates the road map of this research work.

1.4 Research contributions

The main contributions of this research work can be listed as follows :

– Looking at cognitive radio networks from a reliability point of view to investigate how reliability and quality of service can be defined in these networks, and how different cognitive radio features can be leveraged to improve the reliability and quality of service ;
– A novel reliability performance analysis in cognitive radio networks ;
– A general queueing model with interruptions applicable to different CR scenarios and network models ;
– Novel priority queueing disciplines for traffic differentiation and analytical analysis ;
– Analysis and comparison of the baseline buffering (waiting) and switching spectrum access policies in the presence of an unsaturated traffic ;
– A novel greedy and history-aware recovery scheme for cognitive radio networks applicable to different scenarios.

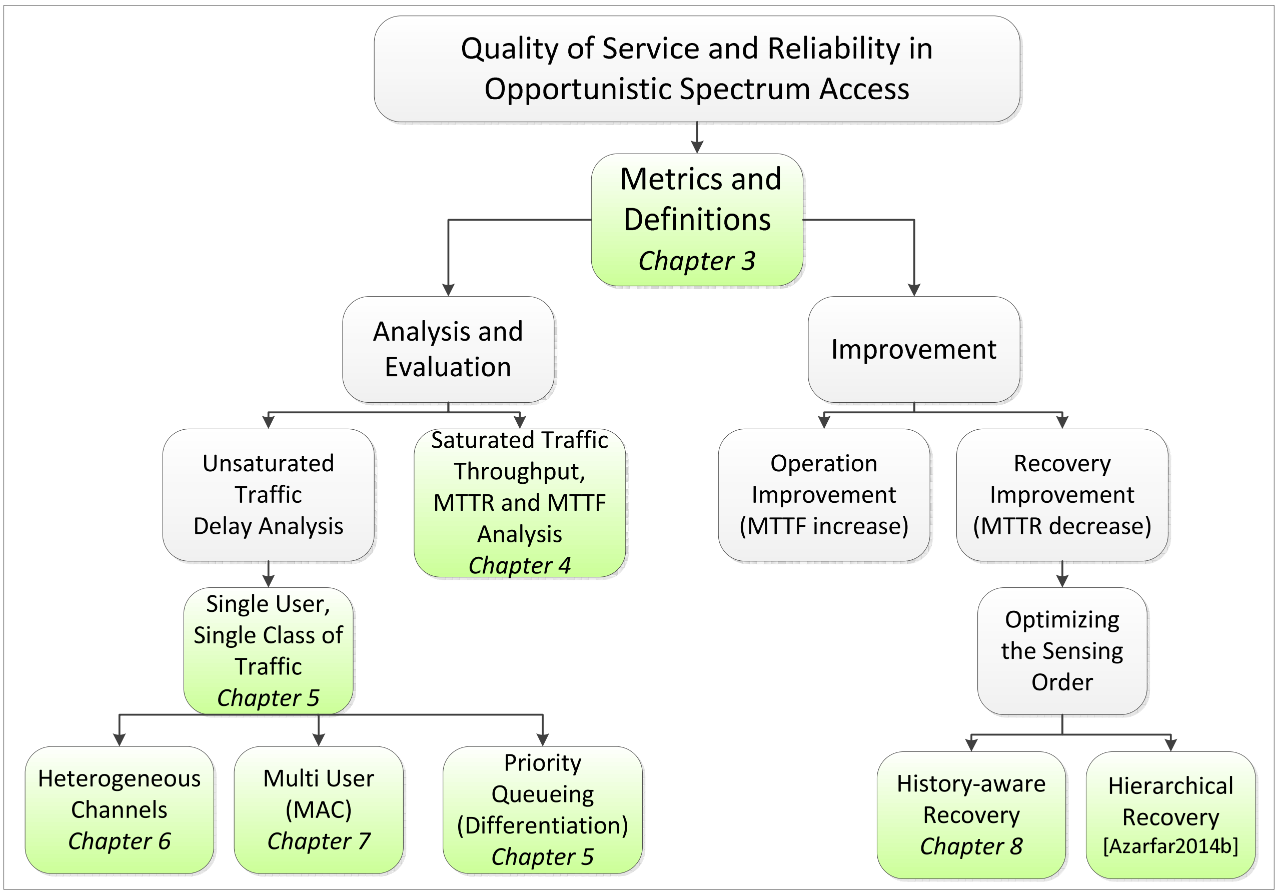

Figure 1.5 The road map of this research work.

1.5 Thesis organization

The organization of the remaining chapters of the thesis is as follows. In Chapter 2, it is discussed how the articles (article-based chapters) provided in Chapters 3 to 8 are related and how they explore different parts of the research subject. The first article in Chapter 3 discusses the basic notions of this research work, such as differentiation, reliability and cognitive radio. In this chapter, it is highlighted how the reliability in a cognitive radio network is defined, and how cognitive radio features can be a means to improve the reliability in wireless communication. The next article in Chapter 4 is a reliability analysis for a cognitive radio network with a saturated traffic. Queueing analysis with unsaturated traffic is provided in Chapter 5, which is the next article chapter. This article also discusses how priority queueing is employed to provide traffic differentiation in cognitive radio networks. Further discussions on the queueing models, how it can be used in networks with heterogeneous channels and how it can be applied to multi-user scenarios, are provided respectively in article Chapters 6 and 7. In Chapter 8, which is the last article chapter, a learning-based approach is followed to improve the recovery performance and thus the reliability in cognitive radio networks. Chapter 9 provides a synthesis and general discussion of the results obtained in this research work, potential applications and extension possibilities, and finally Chapter 10 concludes the thesis with some remarks on future work.

CHAPTER 2

ARTICLE CHAPTERS REVIEW

This doctoral thesis is submitted as a 'thesis by articles' composed of six article chapters, three of them have been published in peer-reviewed journals (3, 4 and 8) and three have been submitted (5, 6 and 7). In this section, we review how those articles explore different areas of the research subject, and how they complete each other to achieve the thesis objectives.

Chapter 3, which was published in *IEEE Surveys and Tutorials* Azarfar *et al.* (2012c), will introduce the basics of the problem, which are the notions of cognitive radio, differentiation and reliability. It is discussed first how reliability is defined in communication networks. The notion of failure is then investigated to review what are the roots of different failure types in wireless networks. Considering the nature of cognitive radio communication networks in which on one side interruptions are frequent, and on the other side they are equipped with notables features, such as spectrum sensing and decision capabilities, it is discussed how any of those CR features may help alleviate the impact of a failure type. This article thus explores how reliability in cognitive radio networks can be redefined, and how cognitive radios can improve the reliability and quality of service of wireless communications.

In this research work and among different features of cognitive radios, we focus mainly on the frequency agility and spectrum switching capabilities of cognitive radio networks. When the current channel has to be vacated either to switch to another channel or to wait for the channel to become available, an interruption occurs. As illustrated in Figure 2.1, considering the nature of opportunistic spectrum access in which interruptions are frequent, we model the event of vacating a channel as a *failure* and consequently the process of spectrum switching as a *recovery* process. In other words, the time spent to sense other channels (if required), select the next channel to operate on that (spectrum decision) and to perform required negotiations and radio alignments is defined as the time to repair or recovery time. With this type of modeling, the objective is to optimize the recovery process especially when different classes of traffic are being served. The thesis can thus be divided into two main parts :

- What is the impact of the recovery process on the performance of a cognitive radio link and network (evaluation and analysis) ?
- How can the recovery be improved (improvement) ?

In the first part and to discuss the impact of the recovery on performance, it is assumed that the recovery process is given, so parameters such as the recovery duration are known. In other words, the goal on those chapters (articles) is not to directly improve the recovery pro-

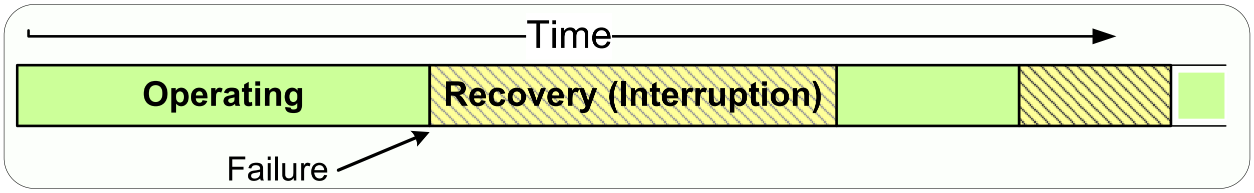

Figure 2.1 General operation model for a cognitive radio link alternating between operating and recovery (interruption) periods.

cess, but to study how the reliability parameters, such as the mean MTTF and availability, and quality of service parameters, such as delay, are affected by the recovery process. The outcome of those chapters are thus mathematical equations and tools besides the insights on the relation of recovery and performance. Naturally, to be able to assess any recovery improvement scheme and to measure how much it has improved the performance, such mathematical tools and equations are required. The next four chapters (Chapters 4 to 7) mostly pursue this goal.

In Chapter 4, which is a paper published in the *Elsevier Journal of Computer Communications* Azarfar *et al.* (2013b), we assume a multi-user multi-channel cognitive radio network and saturated traffic for all users. The impact of different network parameters such as the recovery time, number of users and number of channels on the reliability parameters, MTTF, MTTR and availability is investigated and analyzed in this paper. Using the results of this chapter, we are thus able to provide network design guidelines, such as the optimal number of users and channels when a given MTTF should be met. The other contribution of this chapter is to suggest a failure classification, *hard* and *soft* failures, based on the severity of the failures. Reliability parameters are thus discussed for hard, soft and general failures.

In Chapter 5, which is a paper submitted to the *EURASIP Journal on Wireless Communications and Networking* Azarfar *et al.* (2014d), the impact of recovery in the presence of unsaturated traffic is investigated. The performance metric under investigation is the average system (sojourn) time found from solving an M/G/1 queueing model with interruptions. This is a general model which can be used as a performance evaluation tool in any cognitive radio based network because what is relevant to this queueing model is the distribution of the periods where a cognitive radio node operates and the interruption periods. Therefore, any cause of interruptions can be addressed such as the channel's low quality due to fading, appearance of primary users, interference from another cognitive radio user in a multi-user network, etc. The main insight from the chapter besides the mathematical results is to study how the variation of different parameters, such as the channels' availability and unavailability periods, and the recovery time, affect the average packet delay in CR links. The basic and general model is discussed for a single user network working over homogenous channels.

The proposed queueing model and solution techniques in the previous chapter are then employed first in Chapter 6 to discuss a heterogeneous version of the problem in which a single user operates over multiple channels with different availability and capacity characteristics. This chapter is a paper submitted to the journal of *IET Communications* Azarfar *et al.* (2014e). Regarding the fact that in realistic scenarios, a new channel after an interruption and recovery does not necessarily provide the same service capacity as the channel employed before the interruption, the queueing model should be modified to be able to address such a channel heterogeneity. We provide numerical analysis and analytical approximations for this scenario.

The previous queueing models in Chapters 5 and 6 directly target a single-user cognitive radio network. In Chapter 7, a queueing model is provided for a reference multi-user network with homogenous users and channels. This a paper submitted to the journal of *IEEE Transactions on Mobile Computing* Azarfar *et al.* (2014a) . The interruption periods in this model also include the time spent to compete with other CR users, over a control channel, to reserve a data channel. We are thus able to study the impact of variation of different parameters, such as the number of users and channels, channel access probability (medium access model) and packet length, on the performance of the cognitive radio network. Those results bring an end to the discussions on the performance evaluation in cognitive radio networks (Please see Figure 1.5) when the recovery process is given, and the objective is not necessarily to improve the recovery, but to evaluate its impact.

The first approach to implement differentiated services is provided in Chapter 5. After proposing a general queueing model and solving it for a cognitive radio network with a single class of traffic, the results are then used to propose and solve priority queueing models when multiple classes of traffic are being served. That is, after introducing the general queueing model in the beginning of Chapter 5, the remaining parts of the chapter focus on traffic differentiation in cognitive radio networks using priority queueing. Four different priority queueing disciplines are investigated for a traffic composed of two classes, and for each one, mathematical relations are provided. This part of the chapter thus provides both a performance evaluation discussion (priority queueing models) and a solution to differentiation considering the recovery periods and interruption occurrence.

While in all previous chapters the recovery time is considered to be given, in Chapter 8, we consider a system where the recovery time is the searching time over multiple channels until finding an appropriate channel. In other words, when the channels are sensed one by one, the time until finding a channel which is available and meets the requirements is the recovery time to be optimized. The objective is thus to find the best channel in the shortest time. Improving the recovery is equivalent to shortening the recovery period or/and finding a better channel. For this aim, a history-aware recovery scheme is proposed where the sensing results in each recovery iteration are kept in a database. In the literature, a random sensing order is generally employed which means that the channels are sensed in a random order until finding the first acceptable channel. In this paper, using the information kept in the database, different metrics are defined for the channels and the channels are sensed in the decreasing or increasing order of one of or a combination of the metrics. This chapter is the last article chapter and the article has been published in the *Elsevier Journal of Physical Communication* Azarfar *et al.* (2014c).

At the end of some article chapters, a new section was added and named 'Further discussions (Not a part of the paper)'. Those sections are not a part of the published or submitted papers and are added to provide further discussions or additional information to the article presented in the chapter.

It is also worth noting that in the queueing models, the unit of time is irrelevant and can be any appropriate time unit, e.g., millisecond (ms) or microsecond (us), depending on the network model and other parameters. However, in the simulation parts and unless otherwise mentioned, the time unit is millisecond (ms).

CHAPTER 3

ARTICLE 1 : IMPROVING THE RELIABILITY OF WIRELESS NETWORKS USING COGNITIVE RADIOS

Arash Azarfar, Jean-François Frigon and Brunilde Sansò
Published in IEEE Communications Surveys and Tutorials, March 2011
Department of Electrical Engineering
École Polytechnique de Montréal
C.P. 6079, succ. centre-ville, Montréal, QC, Canada, H3C 3A7
Email : {arash.azarfar, j-f.frigon, brunilde.sanso}@polymtl.ca

To ensure widespread deployment and popularity, next generation wireless services will require a Quality of Service (QoS), and particularly a reliability, that is independent of the radio transmission medium. However, because of the failure-prone nature of wireless networks, providing a reliable communication link and guaranteeing a consistent QoS to users become key issues. In this tutorial, we describe the most common source of failures in wireless networks and provide a systematic failure classification procedure. Drawing from the vast literature on reliability in wireline networks, we then explain how cognitive radios can use their inherent capabilities to implement efficient prevention and recovery mechanisms to combat failures and thereby provide reliable communications and consistent QoS under all circumstances.

keywords Wireless Communications, Failure, Reliability, Cognitive Radio, Prevention, Protection and Restoration.

3.1 Introduction

Wireless services have recently enjoyed tremendous success because users increasingly appreciate the ability to access or share information anywhere and anytime. In return for these conveniences, users have accepted that wireless links are unreliable with inconsistent Quality of Service (QoS) in which problems (dropped or hung connections, variable data rates, delays, etc.) are frequent occurrences. Although users consider these problems as inherent characteristics of wireless networks, as wireless services become more pervasive and replace applications currently provided only in wireline networks, the question of reliability becomes more critical. For example, if a wireless multimedia distribution service is used to replace

AZARFAR, A., FRIGON, J.-F. and SANSO, B. (2011) Improving the reliability of wireless networks using cognitive radios. IEEE Communications Surveys and Tutorials, 14, 338-354.

cable or DSL, users will not accept frequent disruptions during their favorite show. The requirement of reliability and seamless QoS calls for a paradigm change in the design of wireless networks in order for them to offer a reliability comparable to that produced by the fives nines design approach in wireline networks (i.e., 99.999% service availability). The problem is that, because of their nature, wireless networks are inherently more unreliable and error-prone. In addition to the basic hazards, including natural disasters, power failures and hardware failures, that threaten all communication networks, several other factors, such as the random nature of the communication channel and the presence of interferers, affect the reliability in wireless networks. The study of wireless networks reliability is thus a critical aspect for designing network architectures suitable for next-generation wireless services.

A mobile-user connection usually consists of a concatenation of fixed and mobile networks. Any consideration of reliability must consider the entire end-to-end connection. These notions have long been important areas of research in wireline networks Nojo et Watanabe (1993); Ball *et al.* (1995); Vasseur *et al.* (2004) and in wireless network infrastructure Snow *et al.* (2000); Tipper *et al.* (2002). However, the end-to-end reliability is limited by its weakest components. Traditionally, the wireless access link has been perceived as this weakest component and numerous techniques, such as channel coding and diversity, have been proposed at the physical layer to improve the radio-link quality Tse et Viswanath (2005). The design of dependable wireless access networks using a mesh topology has also been a topic of growing interest Akyildiz *et al.* (2005). However, providing dependable wireless access requires a systematic approach that accounts for the different failure modes of wireless links, the available prevention and recovery mechanisms to combat these failures and the relevant performance metrics.

The objective of this tutorial is to discuss the reliability issues associated with wireless networks and explain how Cognitive Radios (CRs) can be used to improve the reliability of wireless networks. In this tutorial, we use the term *reliability* as a qualitative concept which includes all related parameters such as availability and performability and is used interchangeably with robustness or dependability. For the exact definition of these parameters, interested readers can refer to Vasseur *et al.* (2004); Al-Kuwaiti *et al.* (2009). We first present a different perspective on wireless networks based on the notion of reliability, which draws from the vast body of work on the reliability, availability, performability and survivability of fixed wireline networks.

We then discuss how CR technology can be used to implement some of those approaches to build dependable wireless networks. Cognitive Radio is a new paradigm that was introduced in 2000 by J. Mitola to solve the problem of spectrum scarcity and usage inefficiency Mitola III (2000); Haykin (2005); Fette et Fette (2006); Mahmoud (2007), and a new wireless standard based on this technology is currently under development Cordeiro *et al.* (2006); Stevenson *et al.* (2009). Most of the research on CR networks has focused on its spectrum agility features to exploit available spectrum not used by licensed users. However, CR nodes also possess the necessary attributes to make considerable progress in the robustness and dependability of wireless networks Haykin (2005), which has been less explored. Considering the failure classification and failure management approaches that we discuss, the main contribution of this tutorial is in discussing how, for various wireless network failure causes, the different cognitive capabilities of CR nodes can be used to prevent such failures, decrease their occurrence rate and severity or handle the failure after occurrence in a more efficient way. This tutorial provides to the reader the necessary insight to expand this view of cognitive radios in several new directions.

This tutorial is organized as follows. In Section 3.2, we examine the concept of reliability in wireless networks by classifying failures and studying the most common causes of failure. In Section 3.3, we survey the traditional concepts of protection, recovery mechanisms and related performance metrics in wireline networks to study how failures can be prevented or managed after occurrence. The concept of cognitive radio is introduced in Section 3.4 and, in Section 3.5, based on the wireless failure classification and the different failure management approaches discussed previously, we explain how CR features can be used to provide more efficient prevention and recovery mechanisms and build dependable wireless access. Section 3.6 discusses the challenges associated with the development of cognitive radio networks and reviews some of the limiting factors that may slow down the deployment of those networks. Finally, the tutorial is concluded in Section 3.7.

3.2 Failures in Wireless Networks

The challenge in improving the reliability of wireless networks is that the causes and consequences of failures are extremely diverse. For instance, in addition to major failures (such as a base station or mobile malfunction), fading, interference, and battery power, just to name a few, can cause failures. Furthermore, a binary failure model is not adequate. For example, a channel fade might result in a new modulation and coding scheme with a lower data rate. Although the link is still functional, the data rate might be too low to sustain the traffic demand and thus a failure occurs. Therefore, prior to studying prevention and recovery mechanisms for wireless networks, it is important to correctly classify the different failures that can afflict them.

3.2.1 Failure Classification

As illustrated in Fig. 3.1, a classification based on the component type, severity, rate, duration, dimension and scope axes can be used to encompass the most important characteristics of failures in wireless networks. Similar classifications with slightly different failure parameters have been previously proposed Nojo et Watanabe (1993); Vasseur *et al.* (2004) and our classification includes all aspects discussed in the literature. The detection or estimation of these failure parameters can help to devise better prevention methods and aid the recovery mechanisms to select the most appropriate approach, as will be discussed in Section 3.5.

The definition of each failure classification parameter is as follows.

Component type

This parameter indicates the component under failure. In wireless networks, two components can suffer from a failure : the nodes (fix or mobile nodes, base station or spectrum server) or the transmission links.

Severity

Two levels of failure severity can be identified : hard and soft. A hard failure occurs when the communication flow is totally halted. In contrast, a soft failure refers to a situation where the communication flow is not stopped, but the service that can be offered (bandwidth, QoS, etc.) is degraded. This parameter is also known in the litterature as the *Failure Degree* Nojo et Watanabe (1993).

Failure Rate (Frequency)

The failure rate describes the number of times that a failure happens in a specified period. For example, a node failure due to power loss may happen once a month while a failure due to hardware defects happens once every two years.

Duration (Outage Time Nojo et Watanabe (1993))

A wireless network failure can be either permanent or transient. For example, if a user is moving away from a base station, the link failure with this base station will be permanent, while a channel fade causes a transient failure whose duration is determined by the mobile speed.

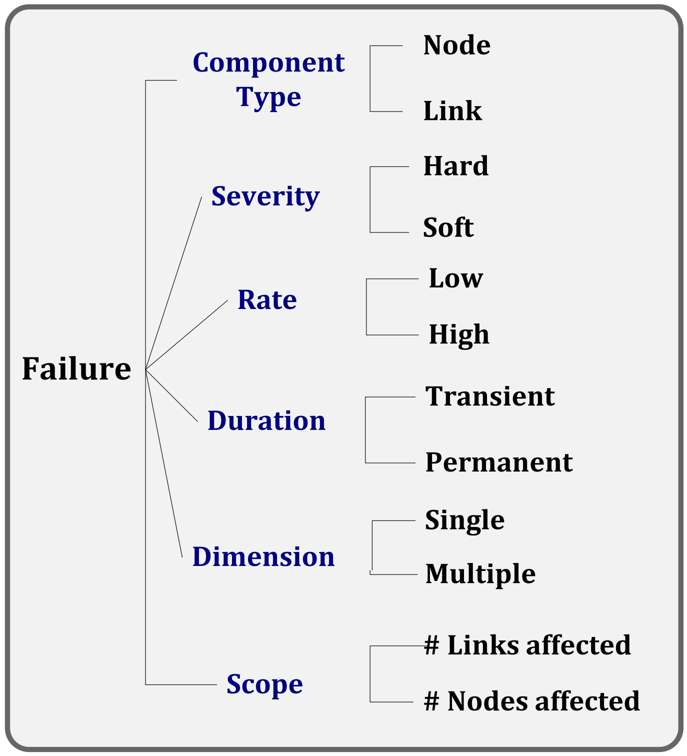

Figure 3.1 Failure classification chart.

Dimension (Failure Cardinality)

The failure dimension indicates whether an event results in single or multiple failures. A single failure dimension implies that, in a short period of time, it is unlikely that multiple failures will occur, whereas a multiple failures dimension indicates that, if a failure occurs, then there is a high probability that other failures will also appear somewhere else in the network. For example, a channel fade has a single failure dimension, whereas the appearance of an interferer has a multiple failures dimension.

Scope (Failure Propagation)

The failure scope is related to the failure propagation concept. That is, a single failure might not only affect the component under failure but also influence the behavior of other components in the surrounding area. The failure scope indicates the area (number of links and nodes) affected by a failure. For example, a link bandwidth degradation in a wireless mesh network might affect the performance of other links in the neighborhood due to the congestion created by the re-routed traffic. However, in a single-hop network, a link degradation only has a local effect on the link.

By the proposed parameters, we try to cover different aspects of failure in wireless networks. The proposed classification is not completely orthogonal and the correlation between the parameters depends on the other factors such as network topology, type of redundancy and application. For instance, in general, permanent failures are hard failures ; however, this is not always true and depends on other parameters. For example, a channel failure due to interference which forces a radio to change its operating frequency is assumed permanent. But, the severity depends on the availability of other channels. If the node quickly finds a new channel, the failure can be assumed soft, otherwise it is a hard failure.

It is also important to consider that a failure can be classified differently depending on the perspective. For instance, in a mesh network, a permanent node failure can be interpreted as a soft failure for other nodes as they are able to change their route. However, for the failed node (user) this failure is hard because it causes the user to get disconnected from the network. Next important point is the perspective and terminology of the failure. As another example, when a protected link is disconnected and the backup link is used, from the link perspective, this event is a soft failure which decreases the overall resource availability in the network. However, this link failure is masked and tolerated at the network operational level. Moreover, the classification of failures also depends on other parameters such as the applications and the specified QoS thresholds like acceptable delay and packet loss ratio.

3.2.2 Failure Causes

Failures in wireless networks occur for various reasons. In this section, we discuss and classify the most common causes of failures and Table 3.1 presents their classification according to the previously proposed criteria.

Node Failure

There are several possible sources of node failure. For example, a power outage, hardware defects and severe software faults are sources of hard node failures because the connectivity is completely lost. In a single-hop wireless network, there is no recovery from such a failure unless a spare node is employed. In a mesh network, the node failure will affect multiple links,

Table 3.1 Classification of most common causes of failure.

Cause of failure	**Compo. Type**	**Severity**	**Rate**	**Duration**	**Dimension**	**Scope**
Node failure	Node	Hard / Soft	Low	Permanent	Single	Several nodes and links
Distance	Link	Soft	Average	Permanent	Single	Limited
Shadowing and fading	Link	Soft	High	Transient / Permanent	Single/ Multiple	Limited
Interference	Link	Hard / Soft	Average	Transient / Permanent	Multiple	Limited
Traffic congestion	Link	Soft	Low	Transient	Single	Limited

but the traffic going through this node may be re-routed. Note that, in both cases, all traffic originating from or targeting the failed node will be lost.

Several types of backup resources may be used (if no backup resource is available, the node failure is permanent). For example, multiple antennas or transceivers can be used. If one component fails (partial failure), the communication link can still use the other antennas or transceivers. However, this technique might result in a lower data rate or link reliability after the recovery. A backup power resource (for example, a battery) can also be used to cope with a main power outage. To preserve its energy, the failure recovery algorithm might elect to reduce the transmit power such that transmission is now only possible with closer neighbors or at a lower data rate. A similar situation can also occur for mobile nodes when the battery level goes below a threshold.

A node failure is permanent and, depending on the availability of backup resources, it can be either a hard or a soft failure. Possible redundancies and the quality of hardware components are such that the rate of node failure is normally low. A node failure will affect a variable number of surrounding links and nodes depending on the network topology. However, it is unlikely that multiple node failures will occur simultaneously.

Link Failure

Path loss, shadowing, multipath fading and interference are the major wireless channel impairments that can cause link failures. A wireless link completely fails when the performance metrics (bit error rate, signal-to-noise ratio, throughput, etc.) are not acceptable. However, in most cases, the signal can still be received with degraded metrics. Bit Error Rate (BER) is the most widespread performance metric and link quality indicator in wireless communication. In general, the BER is inversely proportional to the Signal-to-Noise Ratio (SNR) at the receiver but the exact relation depends on the exact modulation scheme and diversity techniques that are used Rappaport (2001). In a high SNR regime we further have that :

$$BER \propto SNR^{-L}(L > 0) \tag{3.1}$$

where L represents the diversity order of the communication system Tse et Viswanath (2005).

Path Loss

In a wireless network, when the distance between the source and the destination of a transmission link increases due to the users' mobility, the received signal power decreases thereby increasing the BER and packet loss and degrading the link quality. Let d be the distance between the transmitter and the receiver (assuming a constant noise and interference power), we then have :

$$SNR \propto \frac{1}{d^n} \tag{3.2}$$

where n represents the path loss exponent which depends on the characteristics of the environment. In urban areas, n is generally between three and four Rappaport (2001). Because the distance varies gradually, the failure caused by distance is a soft failure, but it can become a hard failure as the nodes become farther apart. The failure due to distance is considered permanent because it cannot be assumed that the nodes will come closer in the future.

Environment Effects (Shadowing and Fading)

Stochastic signal variations, such as shadowing and multipath fading, usually cause transient soft failures. For example, signal degradation due to a building shadow will disappear when the user moves away and small-scale fading causes large signal variation with a displacement on the order of the wavelength. Estimating the duration of those failures can help in implementing efficient recovery mechanisms. These variations decrease the power of the received signal which in turn increase the BER Rappaport (2001).

Interference

In a wireless environment, several users can simultaneously transmit on the same channel, which can create interferences. The SNR at the receiver is proportional to the inverse of the interference :

$$SNR = \frac{P_r}{N + I} \tag{3.3}$$

where P_r is the power of the received signal, N represents the power of the noise and I stands for the total interference. Higher interference thus directly increases the BER of the link. Some technologies, such as spread-spectrum communications, are more immune to interference than others (such as narrowband systems). Therefore, depending on the communication technique, the impact of an interferer can vary from a soft failure to a total link failureTse et Viswanath (2005); Rappaport (2001). In addition, the failure duration depends on the nature of the interferer and can be transient or permanent. For example, a cordless phone will create interference on a wireless network during the time of a conversation but, if a neighbor sets up his wireless network on the same frequency channel, the failure will be permanent. Furthermore, due to the broadcast nature of wireless media, an interferer will usually simultaneously trigger failures on several links.

Congestion

In wireline networks, a high volume of traffic can generate packet loss and delays that can cause severe failures in higher layer communication protocols. In a wireless network, similar phenomena can occur. However, because the wireless channel is shared among several users, a source with a large volume of traffic will degrade not only the performance of his link but also that of the other surrounding users. For example, in random-access protocols such as in IEEE 802.11, a node with a large amount of traffic will increase the contention delay (collision probability) of all users in the network Bianchi (2000). Therefore, traffic increase in one node can cause failures somewhere else in the network.

Special care should also be taken when classifying the cause of a failure. For example, when a node has several operational transceivers and one of them experiences a hardware failure (partial node failure as explained earlier), one of the operating links fails and the radio handles this failure by switching to other transceivers. This implies that we can model these types of node failure as a link failure and consider hardware problems as a new cause of link failures for multi-transceiver nodes. However, a failure in a spare transceiver which is not operational represents a degradation of hardware redundancy and reliability and can not be modeled as such as a link failure.

3.3 Traditional Reliability in Communication Networks

Prior to studying the use of cognitive radios to improve the wireless network reliability in Section 3.5, we will review in this section the traditional reliability concept developed over the years for wireline communication networks. Network robustness has been a major driving factor in the design of wireline networks (such as public switched telephone networks (PSTN) and asynchronous transfer mode (ATM) networks) partly due to regulatory requirements and customer expectations. Network robustness implies network reliability, which generally in a communication network is related to the ability to Tipper *et al.* (2002) :

1. Prevent the occurrence of failures ;
2. Solve and recover from failures.

3.3.1 Prevention Mechanisms

Networks use prevention mechanisms to decrease the occurrence or the severity of failures. Most of these approaches are based on the use of dependable hardware and software for the transmission links and nodes. Other solutions such as selecting less-hazardous environments and equipping communication cables with protective covers are also classified as prevention methods.

The objective of a prevention mechanism is to postpone the occurrence of failures. The most appropriate *performance metrics* to evaluate a prevention mechanism are thus the number of failure occurrences in a period, the probability of a failure occurrence and the duration between two consecutive failures known, respectively, as the *Failure Rate*, the *Reliability* and the *Mean Time Between Failures (MTBF)* of the system.

The *Failure Rate* is the frequency at which a failure happens and can be obtained mathematically from the failure probability density function (PDF) or cumulative distribution function (CDF).

There are different accepted definitions for *Reliability*. Qualitatively, the reliability of a system can be associated to the ability of the system to perform its tasks under some performance and timing constraints Snow *et al.* (2000); Al-Kuwaiti *et al.* (2009). Mathematically, the reliability is given by the probability that no failure occurs during a certain period of time Ross (2006). Let $f(t)$ be the failure probability density function, the reliability is then defined as

$$R(t) = 1 - \int_0^t f(x)\,dx \tag{3.4}$$

and represents the probability that no failure occurs between the time zero and t.

The term *Mean Time To Failure (MTTF)* represents the average time between the return of the system to its normal state and the next failure (i.e., average time to failure). Mathematically, MTTF can be defined based on the failure probability density function as follows :

$$MTTF = \int_0^\infty t f(t)\,dt = \int_0^\infty R(t)\,dt \tag{3.5}$$

To represent the time between two consecutive failures, the term *Mean Time Between Failures (MTBF)* is often used instead of MTTF and is given by :

$$MTBF = MTTF + MTTR, \tag{3.6}$$

where *Mean Time To Repair (MTTR)* stands for the average repair time after a failure occurred. For further definitions, interested readers can refer to (Ross, 2006, Chap. 9). Advanced readers can also find more mathematical details in BARLOW *et al.* (1965).

These different metrics revolve around the same concept and may be used interchangeably in different networks. In this tutorial, we will show in Section 3.5 how the cognitive radio features can be employed to decrease the probability of failure occurrence or equivalently to increase the MTTF of a wireless network.

3.3.2 Recovery Mechanisms

Recovery mechanisms are divided in *Protection* and *Restoration* methods Haider et Harris (2007); Cholda *et al.* (2007). Protection mechanisms are network design and capacity allocation techniques Tipper *et al.* (2002) which assign backup resources in advance, whereas restoration methods attempt to find a solution after the occurrence of a failure. Usually, recovery mechanisms are hybrid and use a mixture of the two approaches. The ability of a network to recover from failures is also studied in conjunction with other concepts such as *Survivability*, *Fault-Tolerance* and *Healing*. For example, survivability is a qualitative concept and is defined as the capability of the system to continue performing its specified tasks when a failure happens Snow *et al.* (2000). In the other words, survivability discusses how a system handles failures by using recovery mechanisms.

Protection Methods

In general, protection methods specify some reserved (spare, backup) resources that will be used when a failure happens, and these backup resources are substituted for the failed ones. The resource substitution can be done automatically by the network or manually by a network administrator. Redundant resources may also be employed actively or passively Shooman (2002); Siewiorek et Swarz (1998). In active protection, before the failure, the backup resource performs the same tasks or plays the role of a load-balancer and shares the specified tasks with the main resource. Aggregated links are an example of active protection IEEE (2008). In passive protection, the redundant backup resource only monitors the status of the system when the main resource is functional. Blocked (backup) links in a Spanning Tree Protocols (STP) without aggregation are an example of passive protection IEEE (2004).

One of the most popular protection mechanisms is Automatic Protection Switching (APS). In APS, a predefined backup resource is substituted automatically when the main resource fails. The most basic approaches are the 1+1 and 1 :1 APS schemes, where each resource has a separate backup W. Lai (2002). Obviously, the cost of this mechanism is high, but it is simple and provides very good and fast recovery in instances of single failures. M :N is another, often employed, APS scheme where one of the M backup resources is substituted when one of the N resources fails. In this case, signaling is required to perform this switching, which can incur a small delay. Obviously, this scheme cannot handle more than M simultaneous resource failures. APS measures can be deployed at different layers of communication protocol and for various resources. For example, at the physical layer, protection could be applied to optic fibers, time slots, subcarriers or wavelengths. In the IEEE 802.22 standard, backup frequencies can be specified for each link that can be used in case of licensed users' appearance or quality concerns Cordeiro *et al.* (2006). However, a physical layer APS measure is sensitive to hazards, such as a backbone accident, that affect all the links routed together.

In the network layer, and jointly with the physical resources, the APS mechanisms are usually based on the network's connectivity and are more robust to single failures. Ring-based, mesh-based and p-cycles can be classified in these survivability methods. A ring-based protection measure is a simple survivable topology that is based on the ability to transmit information in both ring directions. Mesh-based protection is based on more advanced graph connectivity than ring-based protection and can be used to provide link or path protection W. Lai (2002). In link protection (local protection), each link has backup links or paths (paths involve multiple links). If the backup resources are dedicated to the protection of a link, we have disjoint (dedicated) link protection (similar to 1 :1), and if the resources are used to simultaneously protect several links, we have shared link protection W. Lai (2002) (similar to M :N). Path protection (global protection) provides end-to-end protection in the network. It is thus more efficient but more complex than link protection. We can again have either dedicated or shared path protection. P-cycle protection Grover (2003) uses the idea of mixing rings and mesh protection mechanisms. P-cycles provide good protection and low recovery delay similar to ring mechanisms, but their required redundancy is similar to mesh protection Grover (2003). These concepts can be applied to ad-hoc or mesh wireless networks, especially when wireless nodes are equipped with more than one transceivers.

Multiprotocol Label Switching (MPLS) Fast Reroute (FRR) is another local protection approach which is applied to Label Switched Paths (LSPs) et al. (2005). A LSP (1 :1) or a group of LSPs (1 :N) passing through a node is protected by a backup LSP. When this node detects the failure in the main LSP, it activates and uses the backup one. For more details about different protection mechanisms, readers can refer to Grover (2003); Haider et Harris (2007).

Restoration Methods

In restoration methods, when an active resource fails, there is no pre-assigned backup resource and the substitute resources should be found dynamically in reaction to the failure occurrence. Depending on the resource type and operation layer, different schemes can be used. For example, at the routing layer in mesh-based or ad-hoc networks, restoration can be applied at the link or path level W. Lai (2002) :

- Link Restoration : when a link fails, the nodes at both link ends dynamically find a new path to locally re-route the information around the failed link ;
- Path Restoration : when a path fails (due to one or more link or node failures), both ends of the path dynamically find a new end-to-end route around the failures.

In some cases, the recovery methods are hybrid and include both protection and restoration approaches. For instance, in a M :N path protection method, if more than M failures happen (protection fails), a restoration mechanism like re-routing is used to find a new path. Most of these hybrid approaches are cross-layer (multi-layer) healing mechanisms where protection is used for some resources in one layer (usually lower layers) and restoration is employed in the upper layers when the protection fails. For example, a backup frequency may be assigned to a wireless link. When the main frequency fails due to high fading or interference, the user switches to the backup channel. However, if the user finds the backup channel also occupied or unusable, the node requests the base station a vacant channel (infra-structured topology) or routing algorithms are activated to find a new path to the destination (mesh or ad-hoc topologies). An example of such an implementation of cross-layer hybrid approaches in wireless networks can be found in Kant et Chen (2005) where the authors propose protection channels in the physical layer and restoration at the transport layer. For more information about multi-layer survivability, interested readers are referred to Demeester *et al.* (1999).

Performance Metrics

Whereas the recovery mechanisms directly affect the network reliability Cholda *et al.* (2009), appropriate performance metrics are required to evaluate and select the most suitable recovery mechanisms for a communication network. Depending on the network topology, applications, performance goals and traffic type, different scenarios might have different requirements and constraints. Various evaluation metrics have thus been defined Snow *et al.* (2000); Cholda *et al.* (2007, 2009); Grover (2003); Li *et al.* (2008); Kant et Chen (2005) and the most common and important parameters are listed in Table 3.2.

The *Recovery Delay*, which is also known as the MTTR or outage duration, is the average

Table 3.2 Most important metrics used to evaluate recovery mechanisms.

Metric	**Description**
Recovery Delay	The delay between failure occurrence and successful recovery
Time to Next Failure	The quality of the recovery method to prevent further failures
Loss	Frame/Packet loss during recovery period
Complexity	Implementation complexity of recovery mechanisms
Cost	Implementation cost of recovery mechanisms
Resource Availability	Measure of available resources after recovery
Failure Masking	How the method prevents failure propagation

time between the occurrence of a failure and the return to the normal network state after recovery. This parameter is a function of the failure detection time and the speed of the recovery algorithm.

The *Time to Next Failure* is related to the recovery method intelligence and accuracy. It characterizes how the recovery method is able to prevent or postpone the occurrence of failures in the future. This means that an intelligent recovery method can also act like a prevention approach. This parameter has a strong relation with other performance metrics depending on the technology and topology of the wireless network. In a long operational duration, similar to prevention mechanisms, this parameter is evaluated by the MTTF that stands for the average time between the return of the system to the normal state and the next failure.

Without considering other impacts of a failure such as degraded performance or partial failures as described in the resource availability parameter (see below for its definition), the network either performs recovery and is in the recovery/repair state with an average duration equal to MTTR, or is functional with an average duration equal to MTTF. The *availability* metric measures the network's ability to perform its designated function at any given time and indicates the percentage of the time that the system is functional. Since it is a function of the MTTF and the MTTR as follows :

$$Availability = \frac{MTTF}{MTTF + MTTR}, \tag{3.7}$$

we did not include it as a separate metric in Table 3.2.

The *Loss* metric indicates how many packets may be dropped or lost during the recovery period, (e.g., due to the absence of a operational transmission link while a suitable backup resource is established). The number of packets lost depends on several factors such as the packet-dropping policy and buffer management used during the recovery interval, the recovery delay and the traffic characteristics.

The *Complexity* metric quantifies the complexity of the recovery algorithm implementation. For example, a re-routing algorithm that searches the entire network graph is more computationally intensive than a recovery algorithm that acts locally. It can be interpreted as timing or processing cost.

The *Costs* incurred by the implementation of a recovery mechanism are also an important metric for network operators. Resources that are not used actively in the normal network state and are reserved for recovery are a major contributor to the cost of a recovery mechanism. For example, recovery schemes that reserve a backup resource for each active resource (i.e., 1+1 or 1 :1) are considered the most expensive methods.

The *Resource Availability* metric measures the relationship between the available resources before the failure occurrence and after the recovery. It is a very general but important metric. Recovery policies affect this parameter considerably. For example, whereas an exhaustive search might yield the same resources as before the failure, a recovery mechanism might decide to select a backup transmission link or a path with less resources in order to reduce the delay, loss, cost and complexity. The difference between throughput, bandwidth, quality of service and the number of users that can be served can all be included in this parameter. In telephony networks, usually the blocking rate (number of users that can not be serviced) is compared before and after the failure.

The *Performability* parameter, which takes into account the performance of the system after a failure Sansò et Soumis (1991), is often used to quantify the resource availability metric. Performability discusses what percentage of the designated tasks for a system can be performed after a failure. This parameter is directly related to the resource availability parameter as discussed above but no unique definition can be proposed because it depends on several factors such as the topology and network model, application and traffic model and the technology employed in a communication network Sansò et Soumis (1991); Ball *et al.* (1995). Performability measures can be used in the classification of failures as soft or hard, as discussed in Section 3.2.1.

The *Failure Masking* metric indicates the ability of the recovery mechanism to mask the failure and prevents its propagation. This term can be defined by the number of services or users affected by the failure.

Considering these performance metrics, we will show in Section 3.5 how the efficiency of the recovery mechanisms in a wireless network can be improved (lower MTTR, higher MTTF, less failure propagation and etc.) utilizing the cognitive radio features.

3.4 Cognitive Radio Networks

In this section we will review the different features of cognitive radio, and particularly the ones that will be used in Section 3.5 to improve the reliability of wireless networks. A Cognitive Radio node is an intelligent radio that can observe and learn from the environment and adapts its communication parameters based on this knowledge Mitola III (2000). This technology has been proposed as an extension to Software Defined Radio (SDR) Dillinger *et al.* (2003); Jondral (2005); Ulversoy (2010) where this software architecture provides reconfigurability for CR to adjust its communication parameters based on its knowledge and observations. A collection of CR nodes that organize themselves using these cognitive features is called a Cognitive Radio Network (CRN). CRN is a general concept and includes any wireless network from a point-to-point link between two stationary cognitive nodes to a complex cellular or mobile ad-hoc network.

Nowadays, the most known objective of Cognitive Radio (CR) is to solve the problem of spectrum scarcity and usage inefficiency. From 2000 to 2004, reports by the Federal Communications Commission (FCC) stated that on one side, due to extension of the wireless applications and mobile telephony networks, some parts of the spectrum are over-utilized and new requests for spectrum have dramatically increased (spectrum scarcity), while on the other side, several assigned frequency bands are under-utilized and partially used (spectrum usage inefficiency). To solve these problems, as a sub-method for dynamic spectrum access Zhao et Sadler (2007), the idea of using spectrum holes or spectrum opportunities to deploy new networks was proposed. A spectrum hole is a portion of the spectrum which is licensed to a primary network but is vacant in a specific time and geographic area Haykin (2005) and can be temporarily used by a secondary network. This secondary manner of communication implies that a secondary user is able to use a vacant channel when the primary (licensed) user is not present. In wideband technologies, the "vacant" term means that the interference level experienced by primary users will be acceptable (lower than the required threshold) if a secondary network starts operating in this channel. Cognitive Radio, based on its spectrum-awareness, was selected as the best candidate for implementing this concept. A secondary CRN consists of wireless nodes that are spectrum-aware. Using their spectrum-sensing capabilities, they detect chunks of unused spectrum licensed to primary users (e.g., television channels with no broadcaster in the geographic area) and deploy a secondary CRN in the available spectrum. The IEEE 802.22 standard, which is based on this concept, is currently under development Cordeiro *et al.* (2006); Stevenson *et al.* (2009). Note that this mode of operation adds a new cause of link failure when a primary user appears in a CRN operating channel.

The main requirement for a CR secondary user is to respect the priority of primary users. If the physical layer technology allows the co-existence of primary and secondary users in the same channel (e.g. wideband technologies), the CR node should make sure that it does not exceed a given interference level. Otherwise, the CRN has to vacate the channel immediately, which results in a link failure unless a multi-channel (aggregated) communication technology is used.

However, for the purpose of improving wireless network reliability, we consider in this tutorial the general definition of a CR node Mitola III (2000) which possesses the following cognitive features :

- Spectrum-awareness ;
- Location-awareness ;

– Learning and History-awareness ;
– Adaptability and Reconfigurability ;
– Reasoning and Decision-making.

Using these features, a CR node operates in a cognitive cycle illustrated in Fig. 3.2.

3.4.1 Spectrum-awareness

As mentioned earlier, spectrum-awareness implies that a CR node is able to sense the spectrum to find the spectrum holes and estimate their quality considering the interference and environmental effects. Based on the spectrum sensing results, the transmitter and receiver select a common spectrum hole as their operating channel. Sensing may be done periodically or occasionally to verify if the channel is still vacant of the primary users' activity and/or to verify if the channel quality is acceptable. If one of these conditions is violated, the CR node decides either to change its configuration (e.g., transmission power, modulation and coding scheme) to decrease the interference level and compensate the channel effects or to switch to a new vacant channel.

In the sensing process, the CR node attempts to detect the signal of other users. If the CR node has some information about the interference signal characteristics, sensing can be done more accurately and faster using coherent detection mechanisms (e.g., Matched Filter Detection Akyildiz *et al.* (2006)). Otherwise, the CR node can use Energy Detection mechanisms Akyildiz *et al.* (2006) which is fast but not as accurate since different signal sources in the channel can not be distinguished and the possibility of incorrect decisions increases. A more accurate approach is cyclostationary feature detection where the CR node analyzes the spectrum for a long enough period and detects the other users' modulated signal based on their periodicity Akyildiz *et al.* (2006). Noise signals are not cyclostationary so the CR is able to more easily and accurately differentiate them from modulated signal. The cost that is paid for this accuracy is a longer sensing time Akyildiz *et al.* (2006).

When a CR node senses the spectrum, obstacles, distance and fading conditions may affect its sensing accuracy. In a rich fading or high interference environment, it is possible that a CR node detects an unused channel occupied (false-alarm) or a busy channel vacant (miss-detection). When a false-alarm occurs, the CR node does not use the channel for communication although it is vacant. For a miss-detection, the CR node will either interfere in the communication of other users or suffer from high interference level. When a CR node incorrectly detects the status of one channel, one of its neighbors may sense this channel correctly due to its different position and environment. Therefore, the exchange of sensing information among CR nodes in a neighborhood decreases the possibility of false-alarms or miss-detections. Moreover, depending on the available bandwidth, the sensing of the whole

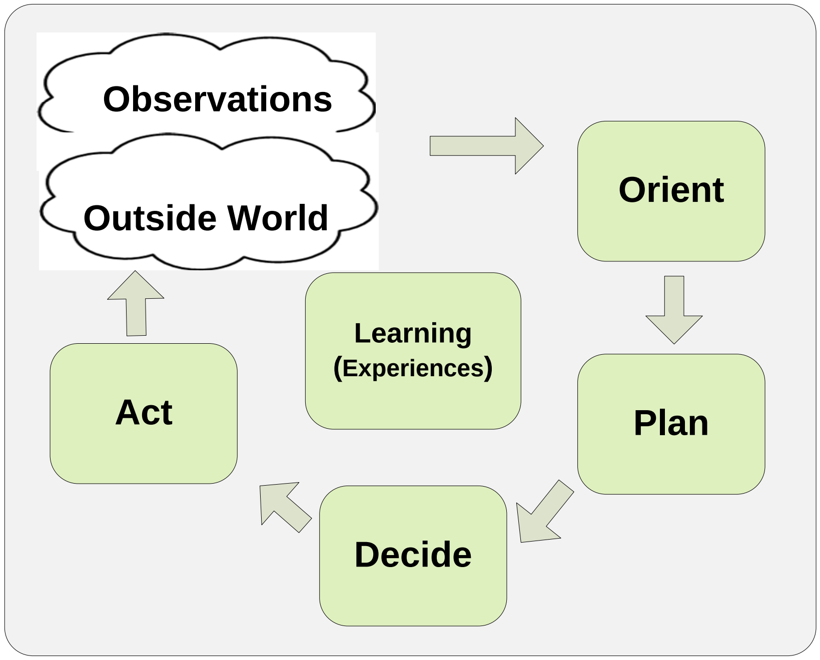

Figure 3.2 Cognitive cycle inspired by the concept proposed in Mitola III (2000).

spectrum for a CR node could be time-consuming. Again, by cooperation of CR users, the task of each CR node can be simplified and the sensing task can be distributed among the nodes, each node senses a narrower band of the spectrum. The distribution of this information among all the nodes makes all of them aware of the spectrum. These methods are called *Cooperative* or *Collaborative Sensing* Cabric *et al.* (2004); Ghasemi et Sousa (2005). In centralized CRNs, instead of exchanging the information with the neighbors in a distributed manner, gathered spectrum information by each user can be sent to the cognitive base station or a spectrum server which analyzes the spectrum, assigns a channel as the next operating channel to each CR node and sends a list of channels to each of them to be sensed in the next sensing period. Although collaborative sensing seems very useful and interesting, several problems may arise. First, synchronization among users, especially in an ad-hoc network, is hard and increases the complexity. Second, the dissemination of the sensing information introduces additional overhead and interference, and needs efficient signaling and routing algorithms. For more details about the different mechanisms of spectrum sensing and related challenges readers can refer to Cabric *et al.* (2004); Akyildiz *et al.* (2006); Ghasemi et Sousa (2008); Yucek et Arslan (2009).

3.4.2 Location-awareness

Location-awareness can greatly help a CR node to have an accurate view of the network. The propagation delay estimation, analysis of the position of the base stations and routing in an ad-hoc network are some of the tasks that can be simplified or improved when a CR node has location-aware mechanisms. CR nodes may employ a Geographical Positioning System (GPS) tool to accomplish this task Fette et Fette (2006); Jondral (2007). When the use of GPS tool is not practical, positioning information may be sent to users by a central entity. In ad-hoc networks, some users can be equipped and send their position to the neighbors. Using directional antennas and distance estimation, other users will be able to find their approximate positions. In (Fette et Fette, 2006, Chap. 8), more information about the usefulness of location-awareness in CRNs and its implementation details is provided. For location-aware routing mechanisms, readers can refer toMauve *et al.* (2001); Stojmenovic (2002).

3.4.3 Learning and History-awareness

In the definition of the CR by Mitola, learning capability implies that a CR node possesses a database which can save the observation results and experienced events in order to use them to take history-aware decision in the future. This capability is implemented as a *Knowledge Base* in some proposed CR architectures and prototypes Mitola III (2000); Jondral (2007); MacKenzie *et al.* (2009). The learning phase is the outcome of observation, planning and decision-making phases. Mitola explains that learning can be very time-consuming and computationally intensive. So, the CR node may have special intervals for learning and in those intervals it does not function as a wireless node. That is, the learning intervals are similar to a sleep period Mitola III (2000).

3.4.4 Adaptability and Reconfigurability

As an extension to SDR, reconfigurability is one of the main features of a CR which makes it enable to adjust different communication parameters based on the current system state. For example, at the physical layer, the frequency, operating bandwidth, modulation and coding scheme, number and configuration of antennas and the transmission power are some of the parameters that can be adjusted. Correspondingly, the sensing process parameters, e.g. the sensing duration or signal power threshold levels, can be adapted.

3.4.5 Reasoning and Decision-making

The huge amount of information that a CR node should process to make a decision and the variety of alternative configurations motivate the need for a reasoning unit in a CR architecture. Just as an example, when a link fails due to the increase of the distance between two nodes, several options are available to them :

- Changing the location if one of the node is mobile ;
- Recovering the link by changing the modulation and coding scheme ;
- Recovering the link by switching to a lower frequency with lower path loss ;
- Finding a new route.

The reasoning unit should analyze all of these possibilities in a very short time to make the best decision. Some performance bounds should be defined and the decision will be made based on these bounds. Spectrum occupancy, SNR, BER, delay, packet loss or a mixture of these parameters may be used as performance bounds.

3.4.6 Cognitive Cycle

All the previously described capabilities are operating inside the framework of the CR cognitive cycle (Fig. 3.2). The cognitive cycle consists of five main stages completed by the learning stage. In the observation stage, the radio senses and identifies the environment to obtain a variety of facts about it. Spectrum awareness and location-awareness methods are part of this stage. During the orientation stage, the CR node adapts its architecture according to the priority and importance of the observed events. Based on the available resources and environmental parameters, the CR creates different plans, decides which plan will be selected and applies the decision by changing the required parameters in various layers. Finally, the CR node can learn from its observations and decisions for future uses.

3.5 Cognitive Radio Networks and Wireless Reliability

Our main objective is to design a wireless system architecture that can counter wireless failures and improve wireless network reliability using approaches similar to those currently in place in wireline networks. As will be discussed in this section, considering its cognitive features and intelligence, the Cognitive Radio has the necessary attributes to achieve this objective. The modified CR cognitive cycle presented in Fig. 3.3 illustrates the inherent capability of CRNs to prevent or recover from failures to improve wireless network reliability. In stage 3, after the environment observation phase and the monitoring of the performance and QoS parameters (stages 1 and 2), the cognitive radio detects whether any new event has occurred or may be occurring in the near future. To make the most appropriate decision, the

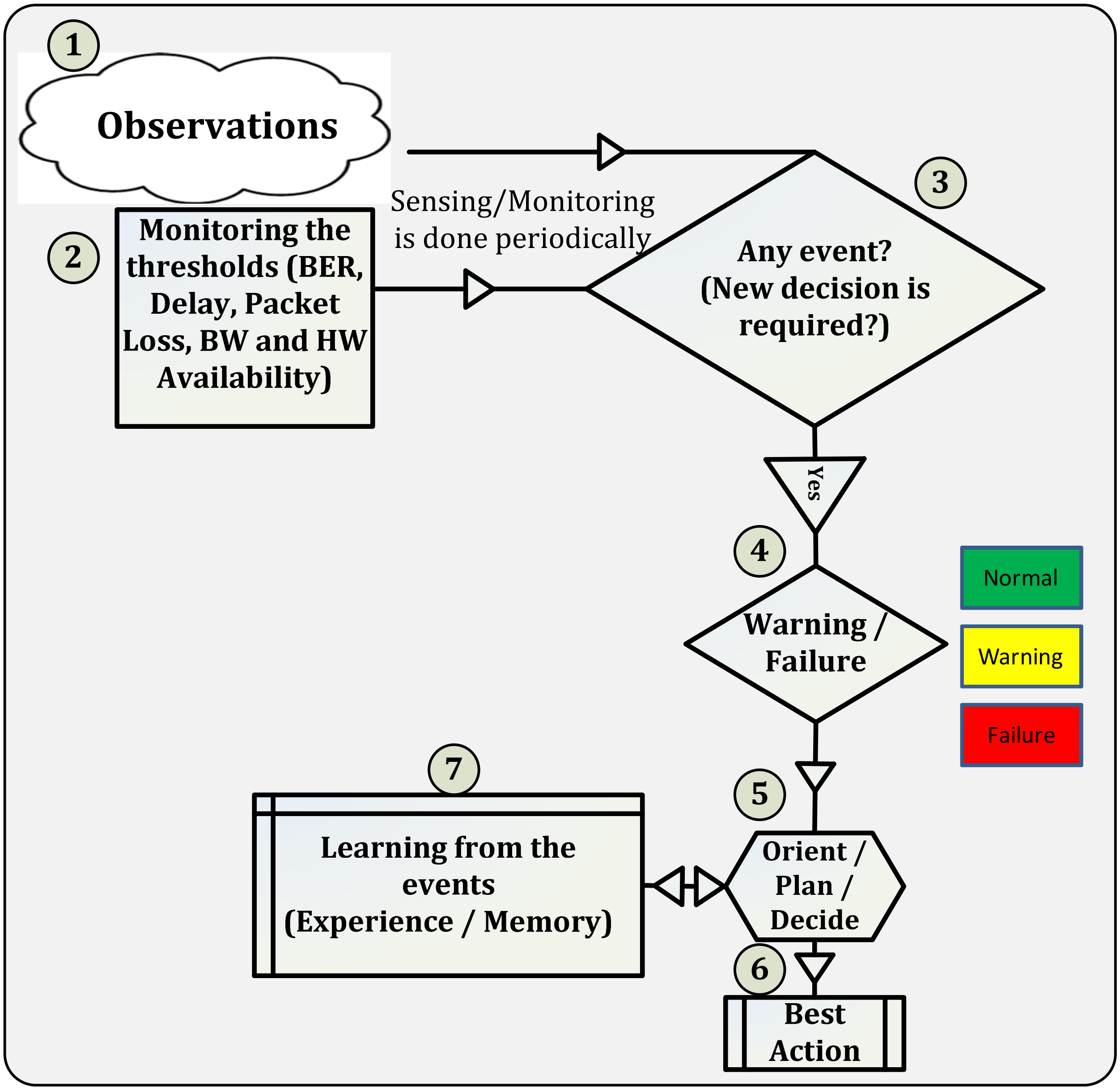

Figure 3.3 Modified cognitive cycle for failure management.

CR node classifies the new event as a Warning or Failure in stage 4. In the former case, the CR deploys failure prevention measures. For example, if a CR mobile station detects that its distance from the base station is increasing, it can switch to a lower modulation and coding to prevent path loss failure. In the later case, the CR node characterizes the failure according to the failure classification chart (Fig. 3.1) and uses the appropriate protection and restoration techniques (stages 5 and 6). The CR node can also learn from the current experiences and observations to help it in the development of more efficient plans in the future (stage 7).

In the following sections, considering the failure causes studied in Section 3.2.2, we discuss how cognitive radios are able to prevent the occurrence of such failures or recover from them using the cognitive features of CRs. Table 3.3 summarizes some of the prevention and recovery methods implemented using cognitive features that can be employed for different failure causes.

Note that the goal of this tutorial is to provide a perspective to the reader on the different possible approaches to use the cognitive radio features to implement prevention and recovery mechanism in wireless networks. Unfortunately, due to the wide array of possibilities, we do not go in details on any of those approaches. However, the material in this section along with the appropriate references provided in this tutorial are sufficient to allow the interested reader to dive further and research or implement the specific CR prevention or recovery mechanism which best suites its needs.

3.5.1 Prevention Methods

Reliable Transmission Techniques

Transmission techniques with higher level of reliability can be employed in CRNs to reduce the probability of link failure (failure rate) or its severity. For example, wideband transmission techniques, such as spread spectrum, frequency hopping and Orthogonal Frequency Division Multiple Access (OFDMA) can be used to increase the wireless network reliability in environments with high levels of interference. Using the reconfigurability and reasoning of CR nodes, when a CRN detects an environment with high level of interference or a primary network using wideband technologies, after the required coordination, it can reconfigure the physical layer to a more appropriate wideband technology. Similarly, transmission parameters such as the channel coding type and rate, the signaling rate and the modulation can be adjusted to increase the reliability of distant users operating with a higher noise level, or to mitigate the impact of interference. In severe fading environments, time, frequency and spatial diversity techniques can be used to increase the system reliability and prevent the occurrence of failures Tse et Viswanath (2005). It is also important to note that, generally, there is a

Table 3.3 Prevention and recovery methods for different failure causes.

Failure Type	**Prevention Methods**	**Recovery (Protection + Restoration) Methods**
Node failure	Components reliability	Re-routing and mesh networking Planning based on HW avl.
Distance	Sensing mechanisms Historical and predefined data	Adaptive transmission techniques Channel switching Re-routing and mesh networking Motion estimation
Shadowing and Fading	Reliable transmission techniques Historical and predefined data Sensing mechanisms	Adaptive transmission techniques Motion estimation Backup channels Channel switching Re-routing and mesh networking
Interference	Sensing mechanisms Reliable channel assignment Reliable transmission techniques Historical and predefined data	Backup channels Channel switching Motion estimation Re-routing and mesh networking
Traffic congestion	Traffic monitoring	Channel aggregation Backup channels Channel switching

tradeoff between reliable transmission techniques that facilitate failure prevention and costs, complexity and throughput.

Sensing Mechanisms

An important feature of cognitive radios is their ability to perform spectrum sensing. A reliable operating frequency channel can therefore be selected based on its interference level (from primary or secondary users) and its attenuation, shadowing and fading characteristics. Increasing the accuracy of the sensing algorithms is thus a primary factor in providing an accurate channel characterization and in improving network reliability. Better sensing algorithms and longer sensing periods can be used to improve the accuracy. In addition, when obstacles, distance, hidden nodes and other special situations limit the sensing abilities of a node, collaborative sensing could be a key in improving the reliability provided by this prevention method.

Historical and Predefined Data

Location-awareness is one of the important features of a CR node Mitola III (2000). Geographical and environmental information can be obtained through a GPS in the CR node, embedded information in packets exchanged between nodes or a central server that sends the most up-to-date global Radio Environment Map (REM) information (Fette et Fette, 2006, Chap. 11). When a CR node knows its location, using learning capabilities, it can record several events (normally not transient) experienced in different locations and times. Fig. 3.4 illustrates this capability. The CR node uses the geographical coordinates of the previously explored areas to remember that, between point 1 and 2, there was a WiFi network. The road between points 2 and 3 is in an urban area and the possibility of interference, shadowing and multipath fading is very high. The area between points 3 and 4 is located near a highway, an airport and a train station. The Doppler spread is therefore higher in this location. After point 4, there was a hilly area and the multipath spread was much larger. When the CR node is approaching these geographical coordinations, a warning alarm is generated that notifies the CR to take adequate measures to prevent the occurrence of failures. For example, the CR node can select a more reliable frequency band or a more robust transmission technique to prevent failures. The CR node or base station may also analyze alternative paths and decides that the path should be modified to reduce the failure probability and thus proposes a more appropriate alternative path. Otherwise, the CR node can activate protection and restoration methods to mitigate the impact of possible failures.

Another example is using previous observation information in the spectrum assignment.

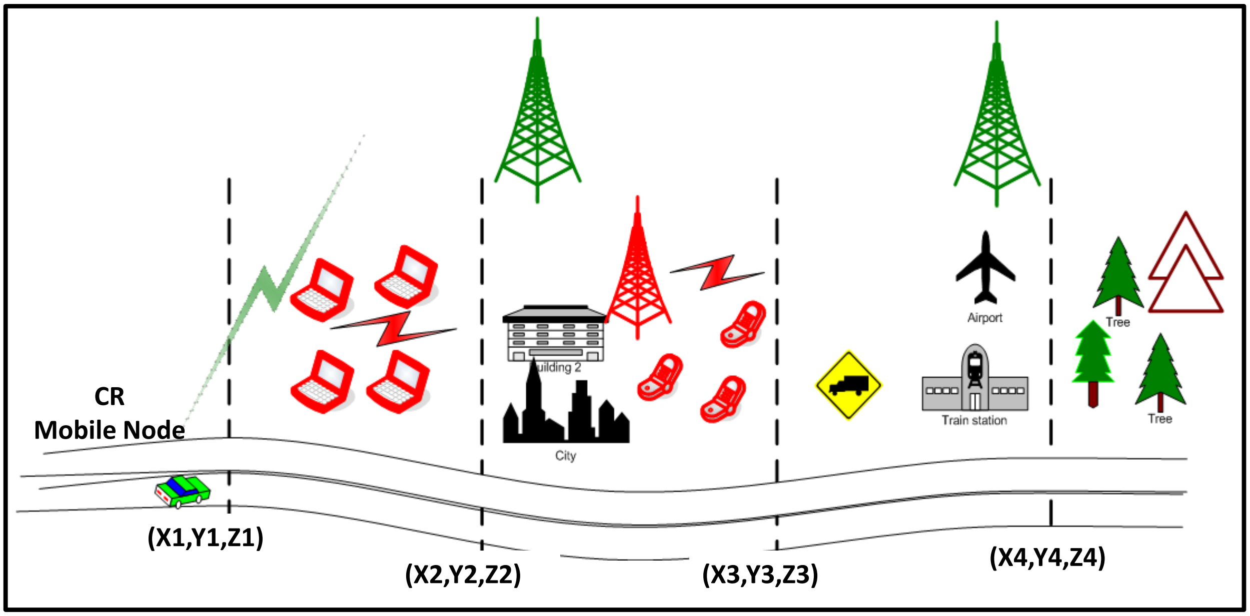

Figure 3.4 Use of predefined and historical data for failure prevention or recovery.

Considering the sensing results of the other channels or previous sensing periods, the CR node is able to prone its decision space when some dependencies between contiguous channels or consecutive time slots exist. For instance, in a network where primary users operate in bonded (aggregated) channels, the CR user may postpone the sensing of channels $N+1$ and $N+2$ when channel N is sensed busy (current slot time) or postpone the sensing of channel N itself when this channel was sensed busy in the previous slot and its probability of return to the available state is low according to its recorded background.

Reliable Channel Assignment

The objective of this prevention method is to select channels that minimize the probability of future failures. For instance, in a multi-channel and aggregated transmission technology, using distributed and non-contiguous channels decreases the probability of simultaneous interference or appearance of primary users in all of the channels Willkomm *et al.* (2005). Sensing and observation results of a long period (statistical information through history-awareness) besides the information that a CRN may obtain through central entities are analyzed to make the best decision for the next operating channel. Some of the proposed CR MAC protocols use statistical data of the channel occupancy to decide the next operating channel more accurately Zhao *et al.* (2007); Chia-Chun Hsu *et al.* (2007). Therefore, even when there is no central entity or negotiation between the neighbors, a CR node does not sense all channels one by one in a random or sequential order. Instead, it uses memorized background information about the channels to create a more efficient sensing order or channel access which decreases the possibility of false-alarms and miss-detections.

In Haykin (2005), the author discusses that cognitive features improve the channel identification tasks such as estimation of channel-state information (CSI) and prediction of channel capacity. This information can be used during the channel selection to decrease the probability of failure in future.

CRNs should also be aware of other neighboring networks. To prevent interference among CR nodes, channel assignment can be done centrally by a server or base station. In an ad-hoc topology where central spectrum management is not possible, collaborative methods can be employed. In this case, spectrum occupancy and available channel information in each node is sent to the other nodes periodically. When a CR node wants to select a new channel, it will consider both its local and its neighbors' information.

In a cellular scenario, the absence of collaboration among neighboring base stations increases the probability of interference noticeably. In this situation, CR base stations may coordinate with each other to establish an efficient static or dynamic channel assignment. The location-awareness features of CR nodes can also be used for a location-aware channel assignment. For example, base stations will not assign the same channels to CR nodes located near the border of neighbor cells.

Traffic Monitoring

A traffic monitoring unit can be employed to periodically measure the status of buffers (level, waiting time, packet loss ratio, etc.), the number of backoffs for random access technologies or other QoS thresholds Kant et Chen (2005). When a node detects that there is a traffic congestion, it can apply congestion-avoidance algorithms e.g. Random Early Detection (RED) to prevent the occurrence of failures caused by congestion. It can also change the frequency band if possible or increase the backoff time to decrease the probability of a dead-lock.

Components Reliability

More reliable hardware and software can be used to improve wireless node reliability. Note that this approach is not specific to CR nodes and is applicable to any wireless system. However, component reliability does not have any impact on link reliability unless for the case that we model a partial node failure (i.e., if one of the multiple operational transceivers fails) as a link failure.

3.5.2 Protection and Restoration Methods

Backup Channels

During the CRN setup phase, the network chooses an operating channel but can also select, in a way similar to APS in wireline networks, a backup channel for protection against failures in the main channel Li *et al.* (2008). This idea is implemented in IEEE 802.22 standard Cordeiro *et al.* (2006). These backup channels can be shared between links if M :N protection is adopted. However, because these channels are devoted to improving the network availability, they are not available for data transmission and the network total throughput is therefore decreased. An important issue is thus to maximize the reliability while minimizing the spectral efficiency loss due to the use of redundant resources.

Channel (Frequency) Switching

If the channel conditions reach a certain threshold where a transmission link failure occurs, the CRN should determine if the failure duration exceeds the acceptable level. In this case, it searches for a new channel and several selection criteria can be used. First, the CRN should switch to a channel with a low interference level and no primary users. It can also, as illustrated in Fig. 3.5.a, use a frequency band with better shadowing and fading characteristics. In the cases where the failure is related to traffic congestion, when the channel bandwidths are not the same, the CRN can search for a wider channel that can accommodate more traffic. This approach is shown in Fig 3.5.b. A channel with better transmission characteristics (e.g., a higher signal-to-noise ratio (SNR)) can also help alleviate the congestion by enabling a higher spectral efficiency.

Some wireless technologies, such as multi-band OFDM, also let wireless nodes communicate using simultaneously several contiguous or non-contiguous channels Mahmoud *et al.* (2009). For instance, IEEE 802.22 provides channel bonding by merging two or three TV channels Cordeiro *et al.* (2006). In case of congestion, the CRN can search for unoccupied subchannels and bond them to its current channels, as shown in Fig. 3.5.c where two subchannels are used to increase the operational bandwidth.

Note that several similarities exist between the channel- switching restoration mechanism and the link-restoration mechanism used in wireline networks.

Adaptive Transmission Techniques

A software defined structure allows a CR to adapt its transmission strategies as a function of the current environment to ensure protection against link failure.

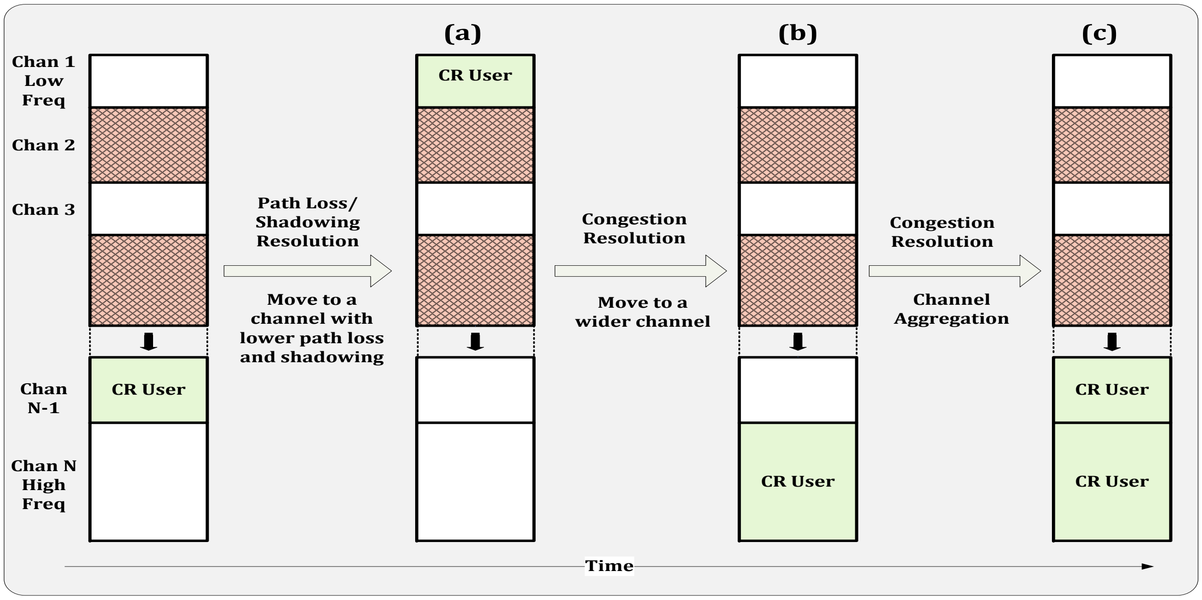

Figure 3.5 (a) Recovery by switching to a frequency channel with lower path loss and shadowing, (b,c) congestion recovery approaches : (b) CR node switches to a wider available channel. (c) CR node occupies two subchannels to increase its available bandwidth.

Following a received SNR variation, Adaptive Power Management can be employed to maintain the link throughput and BER Haykin (2005); Jondral (2007). However, increasing the power should be done while taking into consideration the interference level in the network and the impact of an interference increase on other nodes.

Adaptive Modulation and Coding (AMC) is another mechanism available to cope with SNR variations. AMC adapts the constellation size, the coding scheme and coding rate to achieve a constant BER as a function of the received SNR Rappaport (2001). For example, if there is a link failure due to a SNR decrease, the link can be restored by using a lower constellation size and a lower coding rate. However, the AMC technique affects the link throughput and traffic congestion might then occur.

Another approach to restore the transmission link under fading conditions is to employ *Adaptive Diversity Techniques* Tse et Viswanath (2005). Several parameters of the time and frequency or spatial diversity techniques , e.g. the length of the repetition in a time diversity or the number of parallel channels in a frequency diversity, can be adjusted to the state of the network to provide more robustness following link failures. Adaptive diversity can therefore increase the Time to Next Failure after a link failure at the cost of additional complexity, larger bandwidth or lower throughput.

Other transmitter and receiver internal parameters can also be adapted to the current system state. For example, as a function of the current environment fading characteristics, the equalizer parameters of a narrowband system, the number of fingers in a direct sequence spread spectrum Rake receiver or the number of subcarriers in an OFDM system can be changed and adjusted (Fette et Fette, 2006, Chap. 4). Those changes are more related to the system complexity but do not affect directly other system metrics such as the data throughput. Other adaptive parameters, such as the training and synchronization sequence duration, sensing period or the cyclic prefix length have an impact on the system spectral efficiency. Reconfigurability and adaptability are also present in the higher layers of the communication protocol in CRNs. For example, the spectrum access protocol can be changed to time-slotted following the detection of a cluster/cell with a cluster head or base station.

Re-routing and Mesh Networking

In a multihop network, if a node or a link failure occurs, the information can be transmitted via new routes in the network. In CRN, location-awareness and history of the cognitive nodes help the CR nodes to find a backup path quickly and more efficiently. Location-aware routing protocols have been a research topic in mobile ad-hoc networks (MANET) Mauve *et al.* (2001); Stojmenovic (2002) and can be implemented more easily and efficiently in CR wireless networks due to the adaptive structure of these networks. Thanks to adaptive transmission techniques in CRNs (discussed above), a broken link and consequently a route may be recovered faster and with less complexity by changing the transmission parameters such as frequency, power and modulation scheme without re-routing the broken link locally or changing the end-to-end path globally (link and path restoration mechanisms were discussed in Section 3.3.2).

Just as with wireline networks, the backup routes can be also predetermined or dynamically discovered. However, in the former case, due to the node's mobility, the backup routes must be periodically updated.

It is also important to underscore the difference between the protection offered by backup frequency channels and backup routes. In the first case, if a link failure occurs, the involved nodes select a different operating frequency to reestablish a communication link between them. Additionally, backup channels do not offer protection against node failures. In the second case, the information is re-routed via a new path involving other nodes when a failure occurs. If a link fails between two nodes and other protection and restoration methods are unavailable or inefficient, the CRN can change from a single-hop communication link to a multihop path to maintain the connectivity between these two nodes. Route backups thus offer protection against both node and link failures. However, if the same frequency band is used, the performance after recovery from some causes of link failure, such as interference, is not as good as the one offered by channel protection Li *et al.* (2008). A better approach involves a combination of route and channel backups. A CRN also supports dynamic protection and restoration methods. For example, different routing algorithms may be used for path restoration in a multihop scenario depending on the situation. Re-routing in ad-hoc CRNs and mesh networking are hot research topics in the continuing effort to improve the survivability of cognitive radio networks Chowdhury et Akyildiz (2008); Niyato et Hossain (2009); Akyildiz *et al.* (2009).

Planning Based on Hardware Availability

In CRNs, it is possible to include the current availability status of the hardware parts with the other parameters used to make a new decision. Fig. 3.6 represents a simple case where the availability of backup power (BPWR) is involved in the decision-making. In the normal state, where both the main and redundant resources are available, the CR node has certain functionalities. If a failure occurs, the redundant resource is substituted and the system reliability decreases. The CR node can then decide to change its functionalities, such as its transmission range or relaying in multihop networks, to prevent the occurrence of failures after the recovery. This strategy can also be used by considering the state of the components, such as the battery charge level.

Motion Estimation

A CR node can use an estimate of its trajectory, direction and speed to evaluate a link-failure duration. The failure-duration estimation can also be improved if the CR node has location-awareness capabilities and information about the current environment is available in its knowledge base. Then, considering the performance requirements and available resources, the CR node can decide on the most appropriate protection and restoration method. For example, if the link-failure duration is estimated to be short, adaptive transmission techniques might be a sufficient protection measure. However, if the mobile speed is low and the failure

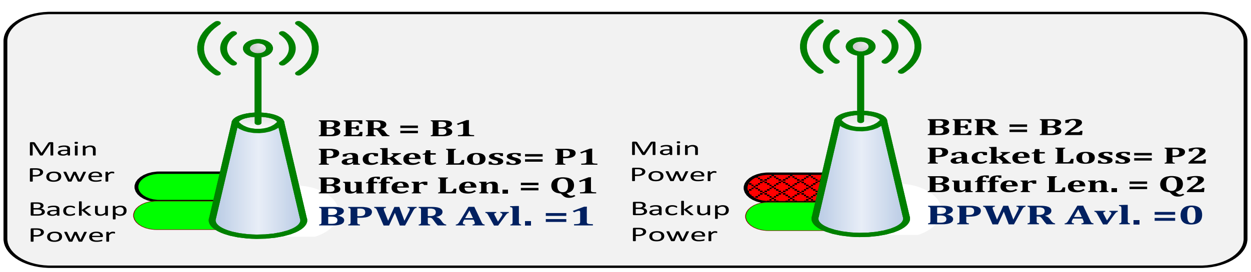

Figure 3.6 Considering the status of the hardware availability to make a new decision.

might last longer, other protective techniques such as using backup channels may be more appropriate.

It should be noted that the cognitive recovery methods we described are not disjoint and may have weak or strong relations based on the topology and type of recovery. For instance, when a CR node misses both its main and backup channel, it uses spectrum sensing and frequency switching capabilities to find a new operating channel. In cognitive routing, the CR node can employ almost all of other recovery methods such as motion estimation, frequency switching and the state of hardware resources to select the best route or routing algorithm.

We can conclude that the cognitive capabilities enable CRNs to use specific recovery methods such as motion estimation and full adaptability that are not available in traditional wireless networks to improve the reliability of wireless networks. It can also provide improvements for existing methods like backup channels. In the latter case, the improvement that can be achieved is quantified and compared using the performance metrics reviewed in Section 3.3.2. For example, a cognitive assignment of backup channels considers the statistical information of the channels (learning and reasoning) and specifies/selects a channel with a low correlation with the main channel to minimize the possibility of concurrently missing both channels. Therefore, compared to the traditional assignment of backup channels, the MTTF (probability of failure) and availability of resources after failure increase, however the MTTR, cost and failure masking are almost the same. Generally, the complexity and hardware cost of the cognitive recovery methods are higher.

3.6 Challenges and Limitations

Although in an ideal scenario a CR node is assumed to be a fully adaptive, reconfigurable and intelligent radio, in implementation, several limitations exist. These limitations have caused that the current deployments and available standards based on the CR technology mainly focus on the spectrum awareness and sensing capabilities of CR. However, with the advance and development of related technological areas such as artificial intelligence, machine learning and digital signal processing, it is expected that other capabilities of CR technology will be developed and implemented in near future Jondral (2007); Mitola (2009). To the best of our knowledge, the limitations to implement an ideal CRN can be categorized in three main areas :

- Decision-making time as an overhead for the main communication ;
- Hardware complexity : cost, processing and power consumption limitations ;
- Channel variations due to secondary manner of communication.

3.6.1 Decision-Making Time

For simplicity, let assume a time-slotted CRN where all activities in the network are synchronized to the boundaries of the fixed slot times with duration T. Then each CR node spends, as illustrated in Fig. 3.7.a, a portion of the slot time, called decision-making time, for observations, spectrum sensing, negotiation and message exchange and recovery (if necessary). The time spent for decision-making is an overhead for the CR node as it is generally not able to use this portion of the slot for transmission. So, a longer decision time implies shorter transmission time or lower throughput. A CR node cannot use this portion for data communication mostly because for sensing mechanisms based on energy detection, the CRN should be silent to be able to accurately detect the presence of the primary users or estimate the level of interference Cabric *et al.* (2004). With feature detection sensing, the necessity of being silent is not as strict, but the sensing time noticeably increases which is not desirable for CR network. Also, when a failure occurs and the CR node decides to change the operating frequency, it should stop the communication in the current operating channel and spend some time to sense other channels to find a new one. This period of time is the recovery delay as discussed in Section 3.3.2 and normally can not be used for data communication.

Although the decision-making time is an overhead, a longer decision period implies higher accuracy and more reliable decisions, which increase the reliability of future communications. Therefore, a considerable part of the research work on CRNs is focused on proposing methods that provide shorter decision-making time with an acceptable level of accuracy and reliability. Particularly, a noticeable part of the decision-making process is spent for spectrum sensing activities : either to monitor the current operating channel to verify the quality, appearance of licensed users and interference level (see Fig. 3.7.b) or when finding a new channel during the recovery period (see Fig. 3.7.c). The sensing time is thus a critical but challenging factor : on one side a longer sensing period increases the accuracy of the sensing and therefore decreases the possibility of false-alarms or miss-detections, which results in higher reliability and throughput(less interference). On the other side, a longer sensing time is equivalent to longer decision-making and recovery delay, which decreases the portion of the slot time available for transmission and therefore degrades the throughput. This implies that there is a trade-off between throughput and reliability and an optimum sensing time should be found which meets the required thresholds and requirements Ghasemi et Sousa (2008); Liang *et al.* (2008); Kim et Shin (2008a).

In addition to the sensing time, the recovery delay also depends on the algorithm that the CRN uses to select the operating channel for the whole network or each CR user known as spectrum decision, spectrum selection or spectrum searching scheme. Normally, the shortest decision time is desirable and the objective in this area of research is to propose faster

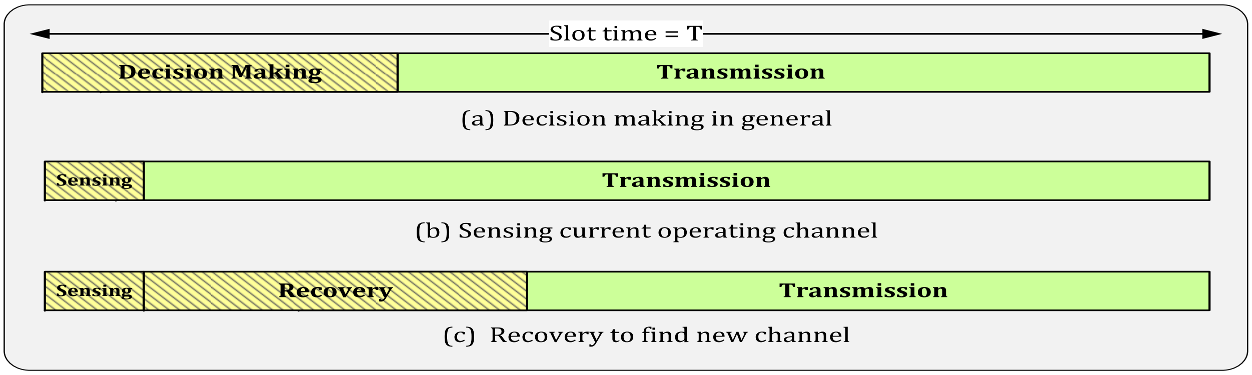

Figure 3.7 Slot structure in a time-slotted CRN. (a) represents a general decision-making time ; (b) and (c) show a specific case of channel switching.

mechanisms Kim et Shin (2008a); Luo et Roy (2007). When the quality of the channels are different, channels may have a quality index (channel characteristics) which is updated centrally by a spectrum server or base station, or in a distributed manner by each CR user itself based on the users' perspective Akyildiz *et al.* (2006). In this case, the best spectrum decision approach is the one that finds in the shortest time the first channel which satisfies the performance and reliability expectations. Also, using historical information of channel occupancy (see Section 3.5.1), the CR node can reduce the decision space and decrease the time spent for finding the next channel.

In conclusion, the decision-making time is an overhead which reduces the resources available for data communications. However, a good decision-making increases the future network reliability and its ability to deliver higher throughput to the user. Also, it is expected that research advances will be able to provide faster decision-making algorithms with good performance.

3.6.2 Complexity and Power Consumption Limitations

Implementing cognitive features such as spectrum sensing, learning and reasoning capabilities considerably increases the hardware complexity, cost and power consumption of the node, which can be a limiting factor in deployment and wide-spread adoption of CRNs. However, recent advances in computer hardware and signal processing as can be seen in PDAs, mobile phones and notebooks demonstrate that the hardware complexity should not be considered as a limiting factor.

Concerning the cost, as explained by Mitola Mitola (2009), it is not necessary for every CR node to possess all cognitive capabilities. A basic architecture with some primary capabilities e.g. spectrum awareness and adaptability can be provided with lower price for the public markets while governmental entities and organizations would use more advanced CR nodes. Furthermore, the ever decreasing cost of technology is an encouraging factor.

For applications such as mobile and wireless sensor networks where power efficiency is critical, giving an important role to CR is more challenging since most of the CR capabilities considerably increase the complexity and processing time and consequently the node power consumption. However, recent advances in related technology fields and the capabilities of the CR networks to employ more power-efficient spectrum management and routing algorithms have provided the possibility of such implementations as has been proposed by Mitola in 2001 Mitola III (2001). Some proposals for CR-based wireless sensor networks Yau *et al.* (2009); Akan *et al.* (2009) and CR mobile cellular networks Sachs *et al.* (2010) have also been proposed in the literature recently.

3.6.3 Channel Variation

Returning to the scenario illustrated in Fig. 3.7.c), a high channel variation implies that in almost all time slots, the CR user performs the recovery and switches to a new channel, which considerably decreases the useful time available for data communication. For example, when a secondary CRN is deployed, although it lessens the problem of spectrum scarcity and usage inefficiency by opportunistically using the unused spectrum licensed to primary users, it adds a new source of failure which is the appearance of primary users in the operating channel which obliges the CR to vacate this band and switch to a new one. The rate and duration of this failure type is dependent on the primary users' activity and traffic type. As the dynamism of the primary users increase, the secondary CRN experiences more failures and spends more time performing link recovery. In Azarfar *et al.* (2010), it is shown that a threshold value for channel variations can be found where using CR technology yields a lower performance than employing a traditional radio with static spectrum assignment and no capability of channel switching. This threshold depends on the rate of channel variations and also the decision-making time. For more examples of the impact of channel variations, readers can refer to Shankar (2007); Pawelczak *et al.* (2008); Su et Zhang (2008); Zhao *et al.* (2007); Huang *et al.* (2008); Zhao *et al.* (2005) where the authors discuss about the performance of cognitive users considering the behaviors of primary networks (channel variations) for different models of medium access.

In Akyildiz *et al.* (2009); Khalife *et al.* (2009), the authors discuss how the routing in multihop CRNs is affected by the channel variations and is consequently more challenging compared to traditional ad-hoc and mesh wireless networks. They show that when the channel variation increases, the efficiency of the existing ad-hoc routing algorithm decreases and the need to propose more specific routing methods for CRNs increases. These appropriate routing algorithms for a CR network, especially for environments with high spectrum variation, should be opportunistic and spectrum-aware Akyildiz *et al.* (2006); Cheng *et al.* (2007); Akyildiz *et al.* (2009); Khalife *et al.* (2009); Ding *et al.* (2010). A spectrum-aware routing algorithm stands for a cross-layer algorithm that finds the route and operating channel jointly. In Opportunistic Routing (OR), a wireless node transmits over any available spectrum opportunity until at least one proper node in the path to the destination receives the packet. OR which is also known as opportunistic forwarding has received considerable attention from researchers in the scope of multihop wireless networks. For further studies, interested readers are referred to Bany Salameh et Krunz (2009); Pelusi *et al.* (2006); Liu *et al.* (2009).

The impact of channel variation explains why licensed TV bands are the most favorite channels for implementation of a CRN Cordeiro *et al.* (2006); Stevenson *et al.* (2009). These bands are usually underutilized and have regular and predictable usage patterns while a WiMAX or WiFi primary network is more dynamic and less predictable.

3.7 Conclusion

We have presented a broad view on failure in wireless networks and network robustness and described a wireless network architecture based on cognitive radios to improve the reliability offered to next generation wireless services. This higher reliability is achieved thanks to the cognitive capabilities of cognitive radio networks that empower these networks to prevent failure occurrence, decrease their severity or recover from failures more efficiently. This approach builds on the abundant literature on reliability in wireline networks and adapts it to the particular context of wireless networks. This tutorial article opens the way for new designs and evaluation approaches for wireless networks with an aim of improving their reliability. However, a detailed investigation of some of the different methods described in this paper and an evaluation of their performance with regards to reliability related metrics is still needed.

3.8 Acknowledgments

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada under Strategic Grant STPG365205. We would also like to thank the Editor-in-Chief and the anonymous reviewers for their instructive and very valuable comments.

CHAPTER 4

ARTICLE 2 : RELIABILITY ANALYSIS OF A CHANNEL RESTORATION MECHANISM FOR OPPORTUNISTIC SPECTRUM ACCESS

Arash Azarfar, Jean-François Frigon and Brunilde Sansò
Published in Elsevier Computer Communications, May 2013
Department of Electrical Engineering
École Polytechnique de Montréal
C.P. 6079, succ. centre-ville, Montréal, QC, Canada, H3C 3A7
Email : {arash.azarfar, j-f.frigon, brunilde.sanso}@polymtl.ca

In this paper, we analyze the promising yet mostly unexplored ability of opportunistic spectrum access (OSA) based on cognitive radios (CR) to provide a robust infrastructure for wireless networks operating in challenging environments with frequent transmission link disruptions. We consider a general network model where the CR users can utilize spectrum sensing and channel switching to determine the status of a channel and use a restoration process when a link failure occurs. We first classify the reliability metrics in CR networks based on the perspective and severity of the failures. We then derive analytical relations for the mean time to failure (MTTF) and mean time to repair (MTTR) of the CR users. With the proposed OSA channel restoration scheme, we show that the MTTF between hard failures, where a user cannot communicate for a long interval, increases exponentially with the number of channels available to the CR users. When a failure occurs, the MTTR also decreases exponentially with the number of channels, thereby providing a highly robust communication environment. Finally, we provide design guidelines that can be used to evaluate the tradeoffs between the number of users and channels versus the required reliability.

keywords Reliability, Mean time to failure (MTTF), Mean time to repair (MTTR), Opportunistic spectrum access (OSA), Cognitive radio (CR), Channel restoration.

4.1 Introduction

Wireless services have enjoyed tremendous success because users increasingly appreciate the ability to access or share information anywhere and anytime. However, providing reliable communications is difficult and challenging due to the error-prone nature of the wireless com-

AZARFAR, A., FRIGON, J.-F. and SANSO, B. (2013) Reliability analysis of a channel restoration mechanism for opportunistic spectrum access. Computer Communications, 36, 947-961.

munication environment. As wireless access becomes more pervasive and replaces wireline access to applications, the problem of providing highly reliable wireless communication services has become critically important. Meanwhile, opportunistic spectrum access (OSA) Zhao et Sadler (2007) using cognitive radios (CR) is a promising solution to the problem of spectrum scarcity and usage inefficiency Akyildiz *et al.* (2006). In addition, due to its cognitive features and reconfigurability, CR technology could also be employed to provide higher reliability or quality of service in wireless communications Mitola III (2000); Haykin (2005); Azarfar *et al.* (2010). These benefits have made CR a technology of choice for next-generation wireless networks, such as in local area (e.g., IEEE 802.11af), metropolitan area (e.g., Cognitive Wimax IEEE 802.16h) and wide area networks (e.g., IEEE 802.22) Granelli *et al.* (2010), and particularly for critical applications that require resilient services, such as emergency and military applications Younis *et al.* (2009).

Despite the importance of analyzing and characterizing the potential benefits of CR to improve the reliability of wireless networks, this topic has not been explored in detail. In Azarfar *et al.* (2012c), we discussed how CR can use their inherent capabilities, particularly spectrum sensing and channel switching Akyildiz *et al.* (2006), to implement efficient recovery mechanisms, to combat failures and, thereby, to provide reliable communications. However, we only introduced the concepts and did not investigate their reliability. Our objective in this paper is thus to analyze the reliability metrics, particularly the mean time to failure (MTTF), mean time to repair (MTTR) and availability BARLOW *et al.* (1965), of a wireless network employing CR nodes with spectrum sensing and channel switching capabilities that enable them to implement a restoration mechanism to find a new available channel in the event of a link failure. To provide a highly robust cognitive radio network (CRN) in challenging environments with frequent link disruptions, we want the MTTF between hard failures, during which a user cannot communicate, to be long, and when a hard failure occurs, we want the MTTR to be short. Note that our work focuses on the link level restoration mechanism and does not consider other strategies that could be employed at higher layers, such as re-routing.

4.1.1 Related Work

Most previous papers have studied the concepts of survivability and connectivity at the network level of wireless networks Snow *et al.* (2000). At the link level, classic wireless communication techniques assume that adaptive channel reconfigurations are not possible on the fly ; therefore, no link level channel recovery is performed. Instead of connectivity and re-routing at the network level or coding and modulation adaptation at the link level, the network model we consider in this paper exploits the possibility of local link recovery by changing the operating channel. Note that compared to frequency hopping, where users follow a predefined order for channel access, the CR users in our work periodically sense the channels and only dynamically change their operating channel to an available channel in the case of a link failure.

Several studies have evaluated the reliability of CRNs that use restoration mechanisms at the routing layer. For instance, Pal (2007) modeled the reliability of a multihop CRN in which the reliability was defined as the reciprocal of the number of transmissions required to successfully deliver a flow in the network. This is a function of two parameters : 1) the average number of nodes in the path, which is related to the classical concept of the shortest path in multihop networks, and 2) the number of attempts (waiting slots) in each hop due to the probability of all channels being unavailable. Therefore, Pal (2007) has considered the blocking probability (waiting), but no discussion has been presented on the different types of failure severity or the MTTF and MTTR.

Furthermore, although several papers have analyzed the effect of parameters such as channel variations, primary users' traffic, sensing time and control channel on the *throughput* of CRNs for different models of channel assignment and MAC protocols Shankar (2007); Zhao *et al.* (2007); Su et Zhang (2008); Azarfar *et al.* (2010), to the best of our knowledge, no work has evaluated the *reliability metrics*, such as blocking probability, MTTF, MTTR and availability, for link layer restoration mechanisms. Indeed, only Li et Qian (2010) has reported a protection mechanism at the link layer for CRNs ; however, the network model that was studied was very specific, and the aforementioned reliability metrics were not analyzed. Additionally, most papers that have analyzed CR systems at the link layer have used the steady-state probability of channel availability instead of considering the transition probabilities. However, as we showed in Azarfar *et al.* (2010), this simplification is not accurate in many situations. Some papers, such as Shankar (2007), have neglected the time that is spent in the spectrum sensing and channel switching processes that are required for restoration. However, as we show in this paper, the restoration time is one of the most important factors affecting the reliability metrics of a CRN.

4.1.2 Contributions and Paper Organization

To the best of our knowledge, this paper is one of the first studies to model failures due to channel occupancy in CRNs and to address reliability metrics such as MTTF and MTTR in a CRN with a link level restoration mechanism. The CRN model that we analyze is based on the opportunistic spectrum access and link level restoration mechanism presented in Section 4.2. This model is similar to the one we described in Azarfar *et al.* (2010). However, in Azarfar *et al.* (2010), we only studied the throughput performance and the blocking probability with simulations. In Section 4.3, we propose a novel reliability model for the CRN based on a channel and user perspective on link failures and the reliability metrics. We also introduce the notion of hard and soft failures to distinguish between different levels of failure severity from the user's perspective. In Section 4.4, we analytically study the probability of blocking, the MTTR, the MTTF and the availability reliability metrics and obtain closed-form expressions. Based on those relations, in Section 4.5, we derive simple expressions for two special cases that clearly illustrate the impact of various parameters on the behavior of the reliability metrics. In particular, we show that in the proposed link level restoration mechanism, the MTTF exponentially increases with the number of available channels, thereby providing a highly robust communication environment. In Section 4.6, we present a simple case study and demonstrate how the relations we obtained in Sections 4.4 and 4.5 can be used as design guidelines for reliable wireless networks based on OSA. Finally, Section 4.7 concludes the paper and provides interesting future research directions for this topic.

4.2 System Model

We consider a system consisting of N non-fading equal bandwidth orthogonal channels. The channels are synchronously operated in a time-slotted fashion as assumed in Zhao *et al.* (2007). Perfect synchronization can be obtained using broadcast messages over all operating channels Mo *et al.* (2008) or over a common control channel (CCC), as assumed in Su et Zhang (2008); Jia *et al.* (2008), or by employing Global Positioning System (GPS) tools as discussed in (Fette et Fette, 2006, Chap. 8). The length of a timeslot is T units of time, where one unit is the time required to send a packet.

As shown in Fig. 4.1, we follow the common assumption in the literature that the availability of the channels can be modeled as identical independent discrete-time Markov chains with transition probabilities α and β Zhao *et al.* (2007); Su et Zhang (2008); Zhou *et al.* (2008); Kim et Shin (2008b). Transmission is possible at the maximum service capacity of the channel when the channel is in the available state, and up to T packets can be transmitted in a slot. No packets can be transmitted while the channel is in the unavailable state. The transition from the available state to the unavailable state is defined as a channel failure (the concept of failure and its impact on the reliability is explored in more detail in Section 4.3) and can be caused, for example, by the appearance of interferers or jamming. We also assume, as in Su et Zhang (2008), that the status of the channel is the same from the perspective of all users. The status of the channel changes only at the ends of the timeslots ; this is also a common assumption in the literature Zhao *et al.* (2007); Su et Zhang (2008).

These N channels are used to create a network of M users (nodes), which are denoted as $U_1, U_2, ..., U_M$. As illustrated in Fig. 4.2.a, the network model could be an infra-structured

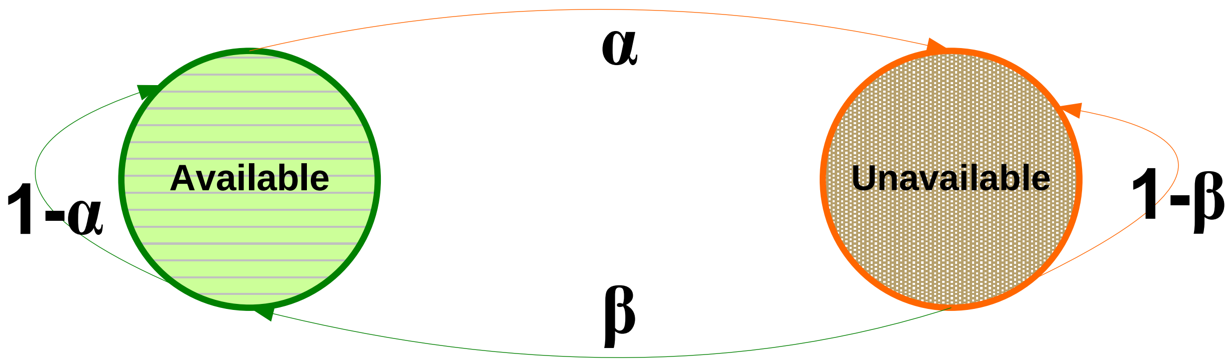

Figure 4.1 Two-state Markov chain channel model.

network in which the CR users communicate to an access point (AP) simultaneously, and each user operates over a disjoint channel. Our analysis is also applicable to an ad-hoc network, as shown in Fig. 4.2.b, where $2M$ synchronized users are scattered in an area, and each pair of users communicates in an ad-hoc manner. A CR user represents a CR transmitter that is communicating with its intended receiver, which can be interpreted as M point-to-point CR links (as in Liu *et al.* (2008)).

Only one user is able to use a specific channel in each timeslot (exclusive access), and the system is non-preemptive (i.e., a user cannot acquire the channel while it is in use by another user). Furthermore, a saturated unidirectional traffic flow model is assumed (as in Su et Zhang (2008)) such that users always have packets that are ready to be transmitted to their destination. Note that the assumption of saturated traffic yields the worst-case reliability metrics.

Fig. 4.3 illustrates the basic opportunistic spectrum access protocol with a *restoration* mechanism Azarfar *et al.* (2010) that we analyze in this paper (the number inside the frame indicates the channel currently being used by the user). At the beginning of a slot, the users are divided into two groups : those who held a channel in the last slot and were able to transmit packets and those who were not assigned a channel and whose transmissions were blocked. The users in the first group will sense their current channel at the beginning of the slot to determine if it is still available (we assume perfect sensing results and discuss the impact of sensing errors later). The initial in-band sensing time T_s is usually short compared to the timeslot ($T_s << T$). If the channel is still available, the user continues to use the same channel and transmits $T - T_s$ packets (this selfishness assumption can violate short-term fairness but is still fair in the long-term ; moreover, the long-term result will be the same if we assume common models from the literature that a user reserves the channel to transmit a long packet flow). Otherwise, a link failure has occurred, and the CR user, along with the

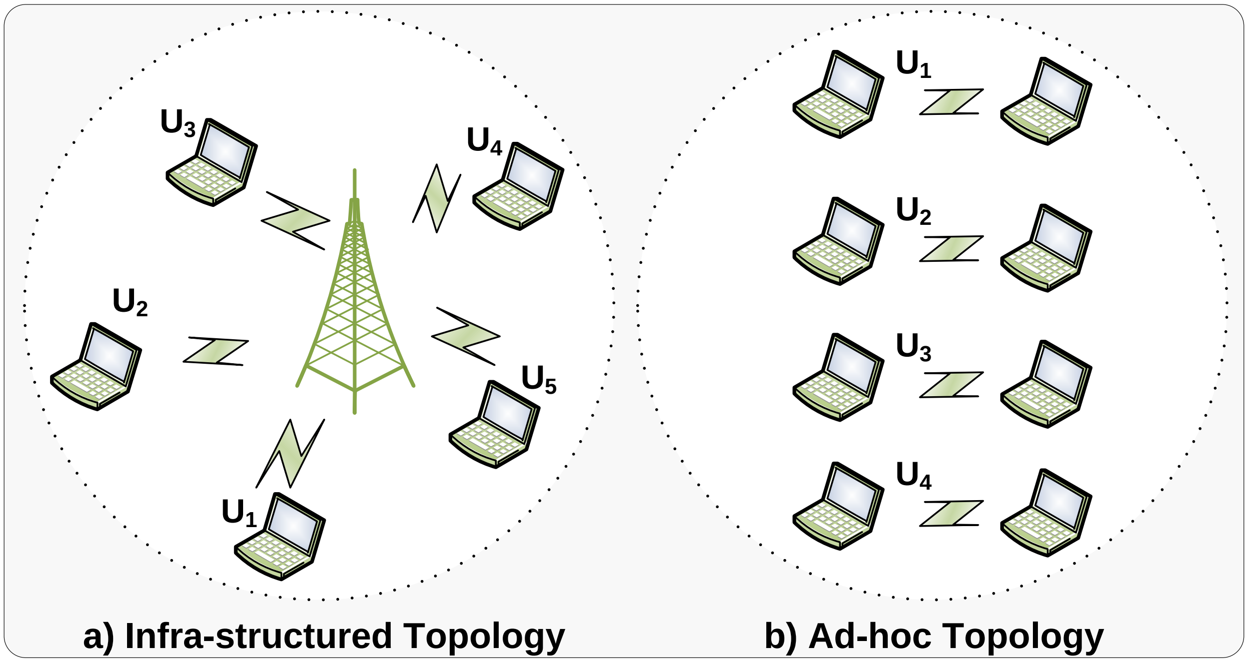

Figure 4.2 Applicable network models.

users who were blocked in the previous slot and the other users who lost their channels in the current slot due to a link failure, starts a contention period and searches the spectrum for a new available channel in the restoration process.

We assume that the network area is small and that all nodes are within the interference regions of the others ; thus, the sensing results of all CR nodes are the same (as assumed in Su et Zhang (2008); Liu *et al.* (2008)). For the theoretical analysis we assume perfect sensing and the impact of sensing errors on the performance is studied via simulations in Section 4.6.2.

In general, the *restoration time* T_r for a user, which is the time required for a user to perform the restoration process simultaneously with other users, is a random variable with a distribution function depending on the sensing time (T_s), the channel transition parameters (α and β), the number of channels (N), the number of users (M), the time required for handshaking to confirm the new operating channel, and the channel selection algorithm Akyildiz *et al.* (2009). Furthermore, some of these parameters can also be correlated. However, the derivation of the recovery time probability distribution is outside the scope of this paper. To make the model analytically tractable, we assumed that the restoration time T_r is constant for all users and all recovery periods. In Section 4.6.1, we then study with simulations the impact of a random T_r on the results. We find that when the constant T_r is chosen equal to the average length of the recovery process, which can be obtained, for instance, by Monte-Carlo simulation of the exact network model, the analytical results are accurate. In the rest of the

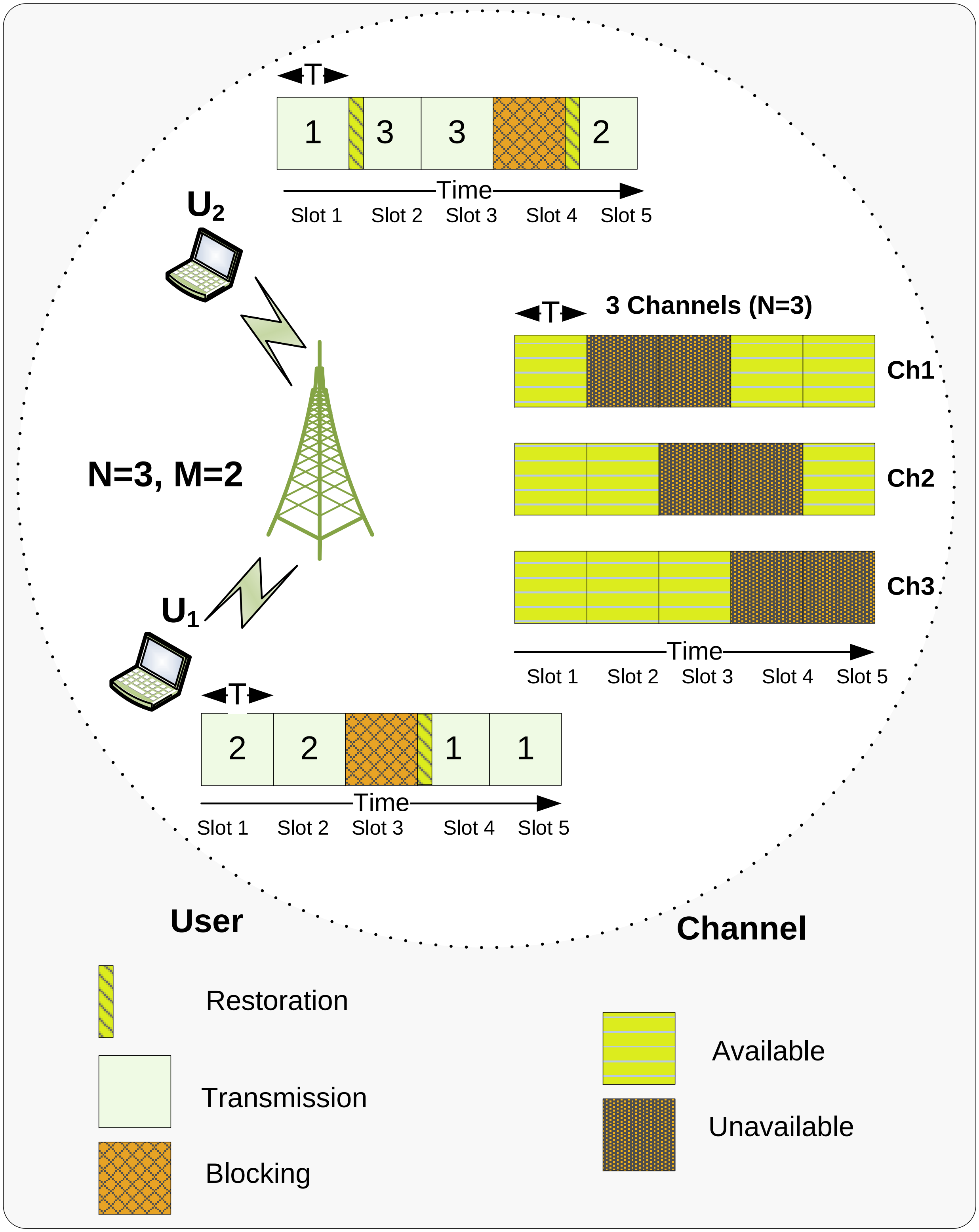

Figure 4.3 Restoration in an OSA CRN with two users and three channels ($N = 3$, $M = 2$).

paper, to keep the notation and analysis simpler we also assumed, without loss of generality, that the restoration time is independent of the different aforementioned parameters. However, the analytical models are still valid if the constant T_r is expressed as a function of the different parameters (i.e., if we replace T_r by $T_r(T_s, \alpha, \beta, \dots)$ where $T_r(T_s, \alpha, \beta, \dots)$ is the average value of the random recovery process with parameters $\{T_s, \alpha, \beta, \dots\}$).

We assume that all the users that enter the restoration process have an equal opportunity to select a channel over the long term. At the end of the restoration process, the CR users who were successful in finding a channel will transmit $T - T_r$ packets in this slot, whereas the unsuccessful users will be blocked and unable to transmit any packets. Note that with the perfect sensing assumption, a CR user can be blocked in a timeslot only if all channels are either unavailable or occupied by other CR users. We will see later that sensing errors may also block a user even when some channels are available. Finally, it is important to emphasize that we do not study the details of a particular restoration mechanism in this paper ; the mechanism must only follow the general guidelines outlined in this section.

4.3 Reliability Modeling

The reliability metrics can be computed based on the channel and the user perspective. The former considers the failures at the channel level due to changes in the channel availability status. On the other hand, the user perspective characterizes the failures that a user experiences during system operation. For instance, the impact of a channel failure on the user's operation might depend on the protection and restoration mechanisms that are employed.

4.3.1 Channel Perspective

From the channel point of view, the MTTF and MTTR reliability metrics only depend on the channel availability model. Considering the Markov model of Fig. 4.1, the number of slots that the channel keeps in the available and the unavailable states follows a geometric distribution with the parameters α and β, respectively. Let $\mathrm{MTTF_C}$ and $\mathrm{MTTR_C}$ denote the MTTF and MTTR of the channel, respectively. We then have that $\mathrm{MTTF_C} = \frac{T}{\alpha}$ and $\mathrm{MTTR_C} = \frac{T}{\beta}$. Note that those would correspond to the MTTF and MTTR of a traditional wireless network that does not use OSA with the ability to switch the channel in the event of a channel failure.

4.3.2 User Perspective

The OSA capabilities of the CR users affect the definition of the reliability metrics from the user's perspective. Indeed, the capability of a CR user to find a new channel and to continue its transmission requires the separation of failures into two cases : one in which the user finds a new channel in a short period of time that can be tolerated by the user (application) and it continues the transmission in the current slot ; and one in which the user is blocked for a longer period of time because it could not find a vacant channel in an acceptable period of time. The first event represents a failure that has a low impact and is acceptable for users, while the second event represents a blocking event with detrimental consequences on the user. The threshold between these events depends on the application's sensitivity to packet loss and delay and the buffering policies. We selected, without loss of generality, one timeslot as the threshold for the occurrence of a blocking event, which implies that the tolerable interruption time for an application is at most one timeslot. Other threshold values could be used and would affect the exact analytical results but not the observed trends. The distinction of tolerable/intolerable recovery (interruption) periods defines two failure events called *Soft* and *Hard* failures, respectively, which are defined as follows :

- **Failure** (General) : If a user holds a channel, a *failure* occurs when the channel becomes unavailable. According to the Markov model of channel occupancy, the probability of this event is equal to α.
- **Soft Failure** : If a failure occurs but the user finds a new channel quickly and is able to continue its transmission, the failure is called a *soft failure*. In Fig. 4.3, the failure that U_2 experiences in the second timeslot is a soft failure. That is, a channel failure occurred at the beginning of the second slot for channel 1, but U_2 was able to restore its communication link on channel 3. Therefore, the user has been quickly reconnected in the same timeslot, and the severity of the failure can be considered as low.
- **Hard Failure** : If a failure occurs and the user does not find a new channel to continue its transmission after the fixed restoration time, the failure is called a *hard failure*. In Fig. 4.3, U_2 experiences a hard failure in the fourth timeslot. That is, a channel failure occurred at the beginning of the fourth slot for channel 3, and because channel 2 is unavailable and channel 1 is used by U_1, U_2 is blocked. In those cases, the user cannot communicate for at least one time slot, and the severity of the failure can be considered to be high.

The reliability metrics can then be defined as follows :

- **MTTR** : The MTTR is the average time spent for restoration, which is the time required for a CR user to find a new channel. The MTTR also corresponds to the average time during which a user is disconnected after a failure and thus should be kept as low as possible.
  - $\mathbf{MTTR^s}$: This is the MTTR for a soft failure, which by definition is equal to one restoration time equal to T_r.

  - $\mathbf{MTTR^h}$: This is the MTTR for a hard failure, which is at least one slot and lasts until the user finds a new channel after a successful restoration process.
- **MTTF** : The MTTF is the average interval length between the time the user starts transmitting after a failure and the occurrence of the next failure. The MTTF should be kept as large as possible for reliable communication.
  - $\mathbf{MTTF^s}$: This is the MTTF between soft failures (because hard and soft failures are distinct events, the time interval between soft failures might include occurrences of hard failures).
  - $\mathbf{MTTF^h}$: This is the MTTF between hard failures (similarly, it might include soft failures).
- **Availability** : The availability is equal to the average portion of the timeslot during which a user is able to communicate. Note that the short in-band sensing time at the beginning of the timeslot, when the user keeps its channel, is also a part of the availability period. The availability A is computed using the following relation :

$$A = \frac{\text{MTTF}}{\text{MTTF} + \text{MTTR}} \tag{4.1}$$

For a saturated traffic model and $T_s << T$, the availability metric is equivalent to the user throughput per timeslot. Because the throughput metric in CR networks has been thoroughly investigated, our focus in this paper is on the MTTR and MTTF reliability metrics.

Another reliability parameter that will be used in the calculation of the MTTR and MTTF is the *probability of blocking* (B), which is the average long-term probability that one user cannot transmit in a timeslot. When a user is blocked in one slot, all the channels are either unavailable or are occupied by other CR users (with no sensing error).

4.4 Reliability Analysis

The objective of this section is to derive analytical expressions for the metrics presented in Section 4.3 for the general CR network model discussed in Section 4.2. We begin in Section 4.4.1 by analyzing the Markov chain model of the CRN. Based on this model, we obtain analytical expressions for the probability of blocking B and the MTTR and MTTF reliability metrics in Section 4.4.2. In Section 4.4.3, we further refine the MTTR and MTTF for the soft and hard failure cases.

These analytical relations are validated using discrete event Matlab simulations that implement the exact system model presented in Section 4.3 for some arbitrary but realistic values of the number of users, number of channels and channel transition probabilities. The restoration time is fixed, and the users participating in the competition have equal chances to obtain an available channel. Due to the homogeneity of the CR users, the simulation results presented here consist of the average of all the users' reliability metrics obtained over five independent simulation runs, where each simulation consists of 70,000 timeslots (because the confidence interval bounds are tight, they are not presented in the figures).

4.4.1 Markov Chain Model

In the literature, a common model to represent an OSA network is to use a Markov chain with 2^N states, where each state $Z = (\phi_1, \phi_2, ..., \phi_N)$, $\phi_n \in \{-1, 0\}$ indicates the status of the N channels. However, because the channels in our model are assumed to be homogeneous, the Markov chain can be simplified by using the number of available channels as the state variable. Each state can be represented by two variables n_1 and n_2, where n_1 is the number of channels that are available in this timeslot and were also available in the previous timeslot and n_2 is the number of channels that have just become available. $n_1 + n_2$ represents the total number of available channels in the current timeslot. Two variables are required because a channel that has just become available can be employed only after a user restoration, and its effective throughput is thus $T - T_r$. It is easy to find the transition and steady-state probabilities of this Markov chain to find an upper bound of network throughput. That is, for $M \geq N$ all available channels are used, and the result would be exact ; however, some channels can be available but idle when $M < N$ and $n_1 + n_2 > M$. The exact throughput would then depend on how many users performed restoration ; this parameter is not available from this Markov chain model. More importantly, the MTTF and MTTR of the users cannot be derived from the Markov model because we would need to track the status of the users. A complete Markov model would need to have states that indicate which of the M users is using each of the N channels or if each channel is idle or unavailable. Deriving transition probabilities and the reliability metrics from this large model would be complicated and cumbersome.

To derive the reliability metrics of the CR users, we thus turn our attention to an approximate model, which will be used in the remainder of this paper. This model is not a true Markov chain because we neglect the dependency of the users, but for large realistic values of N and M, the analytical results are shown to be nearly identical to simulation results that accurately model the CRN. Such approximate models have also been used in the literature, including recently in Wang *et al.* (2012a).

Due to the Markov model of the channels, for a timeslot i, the state of a user in the CRN described in Section 4.2 can be modeled with an approximated three-state Markov chain as illustrated in Fig. 4.4. Transitions between states occur at the beginning of a slot with the transition probabilities indicated in the figure. Our objective in this section is to find

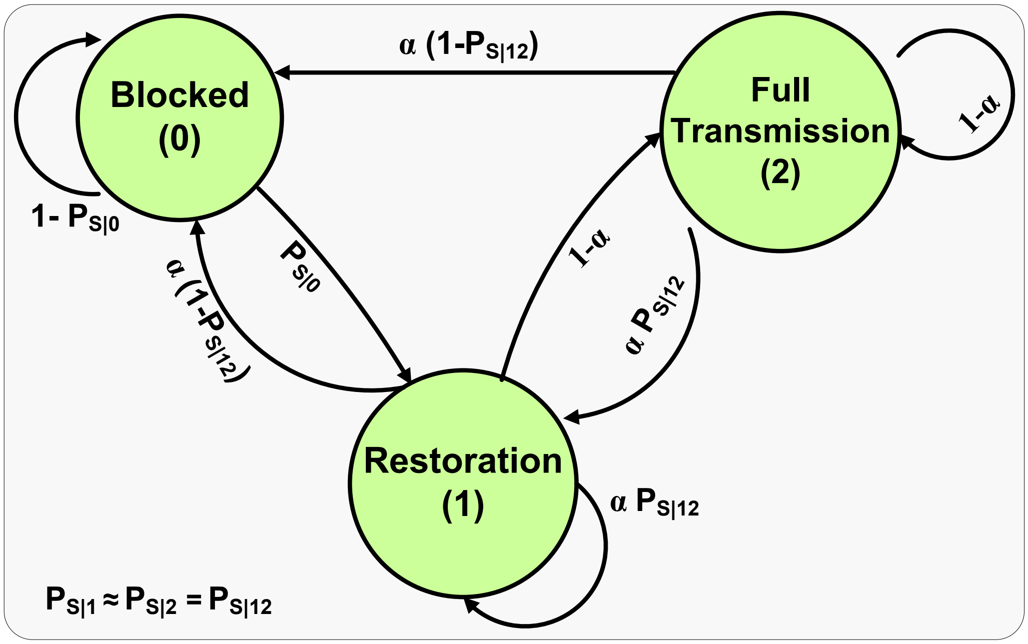

Figure 4.4 Approximate Markov chain model for the state of a CR user.

analytical expressions for those probabilities.

The state of the Markov chain at timeslot i is denoted by $Z_i \in \{0, 1, 2\}$. The definition of the states is as follows :

$Z_i = 0$: In this state, the user could not find any channel for transmission and was blocked. This state is called the blocking state.

$Z_i = 1$: In this state, the user found a channel after a successful restoration process and transmitted $T - T_r$ packets. This state is called the restoration state.

$Z_i = 2$: In this state, the user transmitted $T - T_s$ packets in the same channel that he held in timeslot $i - 1$. This state is called the full transmission state.

Let $S \in \{0, 1\}$ be a random variable for the event of successfully finding a new channel after a restoration process (competition). Given that a user tries to perform channel restoration in slot i, then $P_{S|z} = P(S = 1|Z_{i-1} = z)$ indicates the probability of restoration success given that it was in state z in the previous slot. Inversely, $1 - P_{S|z}$ is the probability of blocking conditioned on the previous state $Z_{i-1} = z$. If the user was in state $Z_{i-1} = 1$, we know that at least one channel was unavailable in timeslot $i - 1$ compared to the state $Z_{i-1} = 2$, where no such knowledge is available. Due to the memory in the channel model, $P_{S|1}$ and $P_{S|2}$ are thus slightly different. However, this difference is negligible as was expected and verified by simulation results (the difference is around 0.01 when N is less than 10 and is much smaller when N is larger). We will thus assume that $P_{S|12} = P_{S|1} \approx P_{S|2}$ and use $P_{S|12}$ to represent $P_{S|1}$ and $P_{S|2}$. When a user was in the restoration or full transmission states in slot $i-1$ and the channel used for transmission has a failure at the start of slot i, we know that at least one channel is not available and that all users attempting restoration will compete for at most $N-1$ channels. On the other hand, when the user was blocked in the previous slot, the number of available channels for restoration in this timeslot can be at most N because all channels can be available. Thus, in general, $P_{S|12} \neq P_{S|0}$. However, for the special case where N is large, the difference between these two probabilities of success can be neglected (the maximum observed difference between $P_{S|12}$ and $P_{S|0}$ was 0.035 when $20 \leq N \leq 30$ and decreases when N becomes larger), and we can use :

$$P_S \approx P_{S|12} \approx P_{S|0} \qquad N \gg 0 \tag{4.2}$$

where P_S is the unconditioned probability of success.

Let $P_B^{\gamma,n,m}$ denote the probability of blocking for a user competing uniformly (as for the uniform random channel access presented in Section 7.2) with $m-1$ other users for n channels, where each channel is independently available with a probability γ. Then, $P_B^{\gamma,n,m}$ is given by :

$$P_B^{\gamma,n,m} = \sum_{n_a=0}^{n} P_B^{n_a,m} f_b^{n,\gamma}(n_a) \tag{4.3}$$

where $P_B^{n_a,m}$ is the probability of blocking for a user competing uniformly with $m-1$ other users for n_a available channels and is given by $\frac{m-n_a}{m}$ if $m \geq n_a$ and 0 otherwise, and $f_b^{n,\gamma}(n_a) = \binom{n}{n_a}\gamma^{n_a}(1-\gamma)^{n-n_a}$ is the binomial probability mass function (PMF), which is the probability of having n_a available channels out of n channels independently available with a probability of γ. We thus obtain :

$$P_B^{\gamma,n,m} = \frac{1}{m} \sum_{n_a=0}^{m-1} (m-n_a) f_b^{n,\gamma}(n_a) \tag{4.4}$$

We now want to derive the analytical relation for $P_{S|0}$, which is the restoration success probability from the point of view of a user who was blocked in the previous slot. Let X be the random variable for the number of users who transmitted in the previous slot and Y be the random variable for the number of users who kept their channel in the current slot. We therefore know that among the N channels, X channels are not available for the competition (i.e., Y channels are kept by their previous owner and $X-Y$ are in failure). Therefore, $M-Y$ users participate in the competition to obtain one of the $N-X$ possible channels. Furthermore, all these $N-X$ channels were unavailable in the previous slot because the user was blocked. Therefore, the probability of availability for the $N-X$ channels is β, and we obtain :

$$P(S=1|Z_i=0, X=x, Y=y) = 1 - P_B^{\beta,N-x,M-y} \tag{4.5}$$

For the X users who transmitted in the previous slot, the probability that each one keeps its channel is $1-\alpha$. Therefore, Y follows a binomial distribution with a PMF $f_b^{x,1-\alpha}(y)$. Furthermore, X can be at most $M-1$ (the user being considered has not transmitted). However, due to the memory nature of the channel, it is difficult to find the exact expression for its distribution. We will thus approximate it with a binomial distribution with a probability of success equal to $\frac{\beta}{\alpha+\beta}$, which is the steady state probability of a channel being available. We then have :

$$P(X=x) \approx \begin{cases} \frac{f_b^{N,\frac{\beta}{\alpha+\beta}}(x)}{F_b^{N,\frac{\beta}{\alpha+\beta}}(M-1)} & x \leq M-1 \\ 0 & x > M-1 \end{cases} \tag{4.6}$$

where $F_b^{N,\frac{\beta}{\alpha+\beta}}(M-1)$ is the cumulative density function (CDF) of a binomial distribution that expresses at most $M-1$ success in N channels, each with a probability of success $\frac{\beta}{\alpha+\beta}$. By unconditioning Eq. (4.5) on X and Y, we obtain :

$$P_{S|0} \approx \sum_{x=0}^{M-1} \left\{ \sum_{y=0}^{x} \left[(1 - P_B^{\beta,N-x,M-y}) f_b^{x,1-\alpha}(y) \right] \cdot \frac{f_b^{N,\frac{\beta}{\alpha+\beta}}(x)}{F_b^{N,\frac{\beta}{\alpha+\beta}}(M-1)} \right\} \tag{4.7}$$

We can derive $P_{S|12}$ in a similar manner. However, in this case, it is possible that no user was blocked in the previous slot and that more than M channels were available in the previous slot (some channels were idle). The user being considered does not know the status of these extra channels in the previous slot. Thus, for the case where $X = M-1$ (X is the number of users other than the user in consideration that transmitted in the previous slot, so it is at most equal to $M-1$), we use the steady state probability, which is equal to $\frac{\beta}{\alpha+\beta}$, rather than considering β in the blocking formula. We then obtain the following expression for the transition probability $P_{S|12}$:

$$\begin{aligned} P_{S|12} \approx & \sum_{x=0}^{M-2} \left\{ \sum_{y=0}^{x} \left[(1 - P_B^{\beta,N-x-1,M-y}) f_b^{x,1-\alpha}(y) \right] \cdot f_b^{N-1,\frac{\beta}{\alpha+\beta}}(x) \right\} + \\ & \sum_{y=0}^{M-1} \left[(1 - P_B^{\frac{\beta}{\alpha+\beta},N-M,M-y}) \cdot f_b^{M-1,1-\alpha}(y) \right] \left(1 - F_b^{N-1,\frac{\beta}{\alpha+\beta}}(M-2) \right) \end{aligned} \tag{4.8}$$

The next step is to find the reliability metrics based on the analysis of the Markov chain.

4.4.2 Reliability Metrics

Probability of Blocking

The probability of blocking B, which is the probability that a user will not be able to transmit in a slot, is given by π_0, the steady state probability of $Z_i = 0$. However, there is an alternative approach to find B. Due to the homogeneity of the users, we can simply consider that we have M users who want to transmit in N channels that are available with a probability $\frac{\beta}{\alpha+\beta}$, the steady state probability of a channel being available. We then obtain :

$$B = P_B^{\frac{\beta}{\alpha+\beta},N,M}. \tag{4.9}$$

Fig. 4.5 compares the probability of blocking as a function of the number of channels obtained using Eq. (4.9) and with simulations. We can observe the accuracy of the analytical model and the linear behavior of $\log(B)$ as a function of N, which will be explored in more detail in Section 4.5.1.

MTTR

As discussed previously, the general MTTR is equal to $T_r + TE[K]$, where $E[\cdot]$ is the expected value of a random variable and K is the number of slots that the user is blocked ($K = 0$ for a soft failure and $K > 0$ for hard failures). K has a geometric behavior, where the probability of $K = k > 0$ is equal to $(1 - P_{S|12})(1 - P_{S|0})^{k-1}P_{S|0}$ and the probability of $K = 0$ is equal to $P_{S|12}$. The general MTTR is then given by :

$$\text{MTTR} = T_r + T\frac{1 - P_{S|12}}{P_{S|0}} \tag{4.10}$$

MTTF

For the MTTF, each instance of time to failure (TTF) is equal to $T - T_r + LT$, where L is the number of slots that the user operates without any failure. This implies that the user's current operating channel remains available for L consecutive slots (remember that for the general MTTF, a failure occurs whenever the user has to perform restoration). Therefore, it has a geometric distribution with the parameter α, and the MTTF of the system is equal to :

$$\text{MTTF} = T - T_r + T(\frac{1-\alpha}{\alpha}) = \frac{T}{\alpha} - T_r \tag{4.11}$$

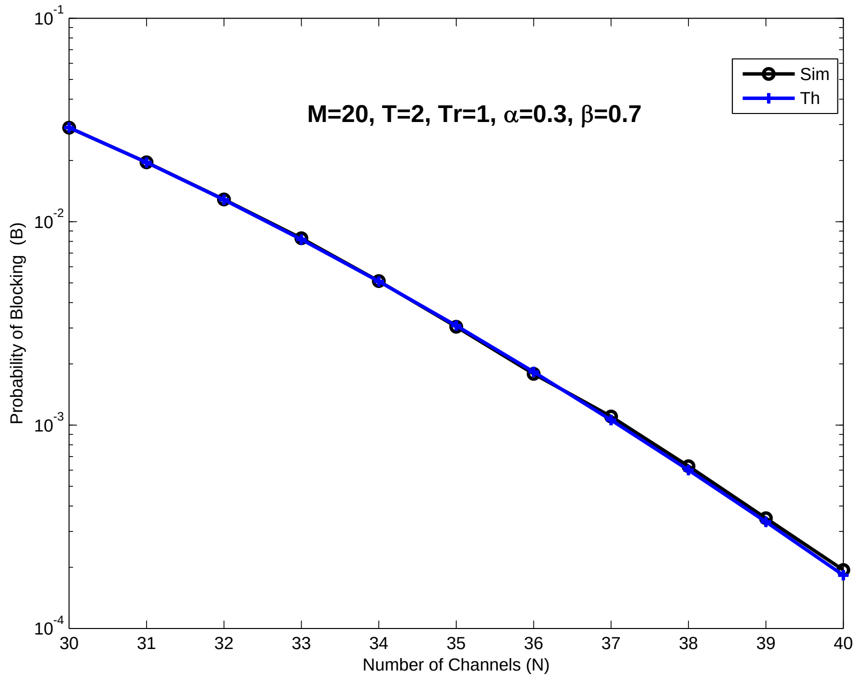

Figure 4.5 Probability of blocking (B) of CR users versus the number of channels (N).

Availability

Based on Eq. (4.1) and the previous results, the availability of the user, which for the case of saturated traffic (if $T - T_s \approx T$) is approximately equal to the average throughput of a user in one slot, is equal to :

$$A = \frac{\frac{T}{\alpha} - T_r}{\frac{T}{\alpha} + T\frac{1-P_{S|12}}{P_{S|0}}} \tag{4.12}$$

Using the best case where $P_{S|12} = P_{S|0} = 1$ (i.e., if a failure occurs, the user always finds a new channel), we can obtain the following upper bound on the availability :

$$A_{max} = 1 - \alpha\frac{T_r}{T} \tag{4.13}$$

Fig. 4.6 compares the availability obtained with Eq. (4.12) versus α with simulation results and confirms the accuracy of the model for both MTTF and MTTR.

It is worth noting that to find the *throughput*, MTTF in Eq. (4.12) should be replaced by $T - T_r + (T - T_s)(\frac{1-\alpha}{\alpha})$ because in each timeslot during an instant of time to failure (TTF), the user transmits at most $T - T_s$ packets.

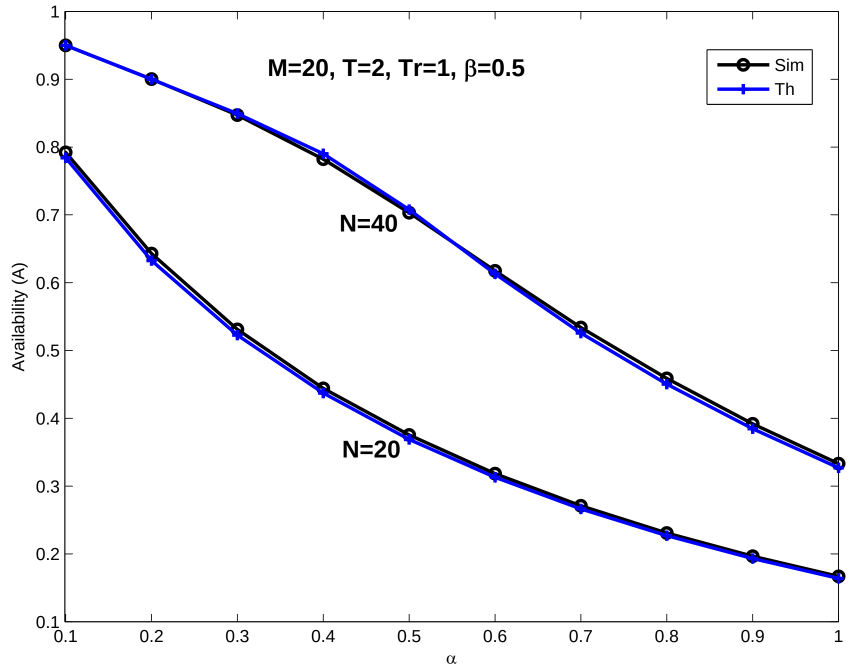

Figure 4.6 Availability of the CR users versus channel variations (α).

4.4.3 Hard and Soft Failures

In this section, separate relations are derived for MTTF and MTTR for the hard and soft failure cases. For soft failures, we have $\mathrm{MTTR}^{\mathrm{s}} = T_r$. For hard failures, $\mathrm{MTTR}^{\mathrm{h}}$ is similar to the general MTTR except that the probability to have a restoration time equal to T_r is zero (the user is blocked for at least one slot). We thus have :

$$\mathrm{MTTR}^{\mathrm{h}} = T_r + \frac{T}{P_{S|0}} \tag{4.14}$$

Fig. 4.7 presents the analytical and simulation results for the general failure MTTR, the soft failure $\mathrm{MTTR}^{\mathrm{s}}$ and the hard failure $\mathrm{MTTR}^{\mathrm{h}}$. As expected, the general MTTR is larger than $\mathrm{MTTR}^{\mathrm{s}}$ and less than $\mathrm{MTTR}^{\mathrm{h}}$. The difference with $\mathrm{MTTR}^{\mathrm{s}}$ increases with α because it results in a higher probability of blocking, which implies that the time to find a new channel increases. For lower values of α, MTTR and $\mathrm{MTTR}^{\mathrm{h}}$ converge to their asymptotic values of T_r and $T + T_r$, respectively. It is worth noting that the small gap between the simulation and analytical results comes from the fact that the derived probabilities of success (Eqs. (4.7) and (4.8)) are approximations. Another interesting point is that the difference between the general and hard MTTR is approximately constant and equal to $T = 2$, which directly follows from the fact that $\mathrm{MTTR}^{\mathrm{h}} - \mathrm{MTTR} \approx T$ when $N \gg 0$.

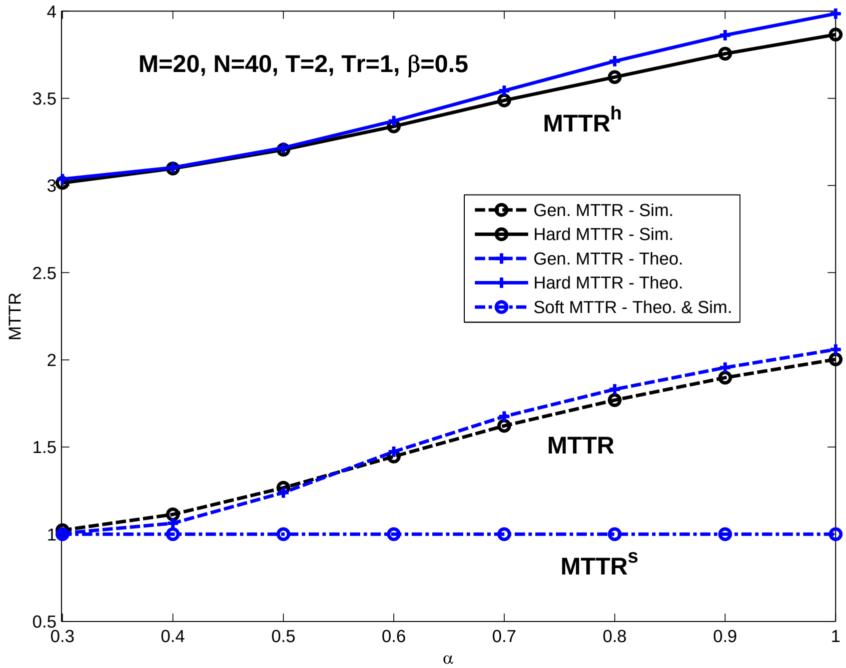

Figure 4.7 Comparison of general, hard and soft MTTRs of a user ($N = 40$, $M = 20$, $T = 2$, $T_r = 1$ and $\beta = 0.5$).

To compute the MTTF for soft failures, we must consider the cases where no failures or only hard failures occur between two soft failures, which means that the user may be blocked several times during this interval. This metric can be obtained easily from the state 1 recurrence time in the Markov chain (Figure 4.4). However, the metric is not useful, and thus, we do not derive an analytical expression for it in this paper.

For the MTTF of hard failures, if there are L slots between the end of the restoration and the next hard failure, only soft failures or no failures occur in the first $L-1$ slots ; they occur with a probability $1 - \alpha + \alpha(P_{S|12})$. A hard failure occurs in the last slot. L thus follows a geometric distribution with the parameter $\alpha(1 - P_{S|12})$, and we obtain :

$$\mathrm{MTTF^h} = \frac{T}{\alpha(1 - P_{S|12})} - T_r \tag{4.15}$$

Fig. 4.8 compares the analytical expressions and simulation results for the general MTTF and hard failure $\mathrm{MTTF^h}$. As expected, $\mathrm{MTTF^h} > \mathrm{MTTF}$, and the simulation results also validate our analytical model. Note that the analytical expressions for both MTTR and MTTF were validated with simulation results for several other parameter values, which are not included here due to space constraints.

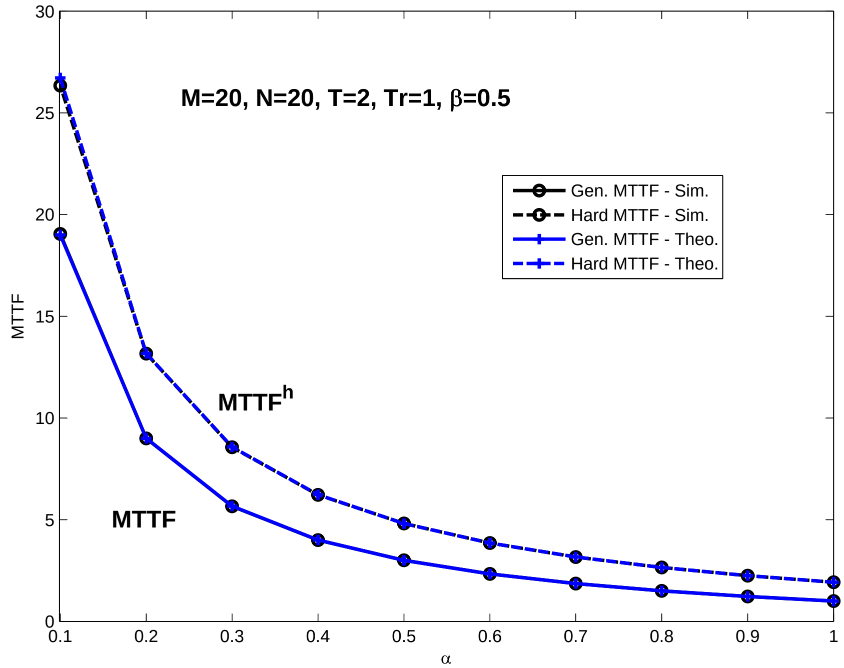

Figure 4.8 Comparison of general and hard MTTF of a CR user ($N = 20$, $M = 20$, $T = 2$, $T_r = 1$ and $\beta = 0.5$).

4.5 Analysis of Asymptotic Cases

We have shown that the analytical relations found for the reliability metrics in the CRN model are valid and can be used to predict the performance of the CRN for different parameters. However, these relations are quite complex and do not provide significant insight into the impact of the various parameters on the reliability metrics. In Sections 4.5.1 and 4.5.2, we analyze two special cases where the network is lightly loaded (i.e., $M \ll N_a$, where $N_a = \frac{N\beta}{\beta+\alpha}$ is the average number of available channels) and loaded (i.e., $M \geq N_a$). These analyses clearly illustrate the impacts of various parameters on the behavior of the reliability metrics.

4.5.1 Analysis of a Lightly Loaded Network ($N_a >> M$)

Considering Eq. (4.9), the probability of blocking can be bounded as follows :

$$B \leq F_b^{N,\frac{\beta}{\alpha+\beta}}(M-1) \tag{4.16}$$

Furthermore, when N is large enough such that $N_a >> M$ and $\frac{\beta}{\alpha+\beta}$ is not close to zero or one, we can approximate the binomial distribution with a normal distribution and have :

$$F_b^{N,\frac{\beta}{\alpha+\beta}}(M-1) \approx Q\left(\frac{M-1-N\frac{\beta}{\alpha+\beta}}{\sqrt{\frac{N\alpha\beta}{(\alpha+\beta)^2}}}\right) \tag{4.17}$$

where $Q(x)$ is the complementary CDF of a normal (Gaussian) distribution. Using the Chernoff bound of $Q(x)$, we then obtain :

$$\ln B \leq \frac{-1}{2}\left[\frac{M-1-N\frac{\beta}{\alpha+\beta}}{\sqrt{\frac{N\alpha\beta}{(\alpha+\beta)^2}}}\right]^2 \tag{4.18}$$

and

$$\lim_{N\to\infty}\frac{\ln B}{N} \leq \frac{-\beta}{2\alpha}. \tag{4.19}$$

This explains the linear relationship of $\log B$ versus N that was observed in Fig. 4.5. Furthermore, this is a powerful result because it indicates that we can make the probability of blocking as small as required by increasing N and that the slope of the decrease in the log scale is controlled only by the parameters of the channel. A similar linear relationship in the log scale with the number of users can also be derived when $M \to 0$.

Because we assume that $N \gg 0$, we can use Eq. (4.2) and the fact that $B = P[Z_i = 0]$ to obtain from the steady-state analysis of the Markov chain that :

$$P_S \approx \frac{(1-B)(\alpha)}{B+(1-B)(\alpha)}. \tag{4.20}$$

Because $N_a >> M$, $B \approx 0$ and therefore :

$$P_S \approx 1 - B/\alpha. \tag{4.21}$$

Substituting into Eqs. (4.14), (4.15) and (4.12), we obtain the following approximation for the reliability metrics in a lightly loaded network :

$$\mathrm{MTTR^h} \approx T_r + T + TB/\alpha \tag{4.22}$$

$$\mathrm{MTTF^h} \approx \frac{T}{B} - (T_r) \tag{4.23}$$

$$A \approx A_{\max}(1-B). \tag{4.24}$$

Using the Chernoff bound, we can see that $\mathrm{MTTR^h}$ and A decrease exponentially toward

$T + T_r$ and $A_{\max}$, respectively, and that $\log \mathrm{MTTF^h}$ is linear for $N \gg 0$ with a slope of $\frac{\beta}{2\alpha}$. The reliability metrics (minimum disconnection time, large intervals between failures and large availability) of the wireless network can thus be improved as required by increasing N up to the limits dictated by the technology in use (frame time and restoration time). It should be noted that contrary to channel variations, the number of channels is a system parameter that is controlled by the CRN designer (i.e., the number of channels that should be leased for opportunistic access from the licensees). Therefore, knowledge about the optimal number of channels to be leased for a CRN to obtain a given required reliability is useful and important.

It is worth mentioning that in a CRN and particularly in an infra-structured network, the users can be aware of the status of other users' channels because this information is broadcast. Therefore, a user knows the exact status of some other channels during the restoration, and the transition probabilities can be used instead of the steady-state probability ($\frac{\beta}{\beta+\alpha}$), which increases the accuracy of the results (we have not used such an assumption). Moreover, when M and N are similar, the user knows the status of many channels ; therefore, only a few remain to be sensed during the restoration (not all N channels), which means that the assumption of a fixed upper bound for the restoration time is realistic.

4.5.2 Analysis of a Loaded Network

In this case, the average behavior of the system can be evaluated using the assumptions that $N_a \leq M$, which implies that the probability of having idle channels is very low and negligible, and $N \gg 0$ such that the approximation in Eq. (4.2) can be used. Because the probability of having unused channels in a slot is low, the probability of blocking can be approximated by $\frac{M-N_a}{M}$, where $N_a = N\frac{\beta}{\alpha+\beta}$ is the average number of available channels. Substituting into Eq. (4.20) (which was obtained from the Markov chain using only the $N \gg 0$ assumption), we then obtain the following expression for the probability of restoration success in a loaded network :

$$P_S \approx \frac{N\alpha\beta}{M(\alpha+\beta) - N\beta + N\alpha\beta}. \tag{4.25}$$

When the network is loaded, the $MTTR$, $MTTR^h$ and $MTTF^h$ of the network can be approximated by simpler relations that are derived by substituting Eq. (4.25) into Eqs. (4.10), (4.14) and (4.15), respectively. For example, we have :

$$\mathrm{MTTR^h} = T_r + T\left(\frac{M(\alpha+\beta)}{N\alpha\beta} - \frac{1}{\alpha} + 1\right). \tag{4.26}$$

4.6 Discussion

Fig. 4.9 and 4.10 show the theoretical $\mathrm{MTTR^h}$ and $\mathrm{MTTF^h}$, respectively, versus N for steady-state probabilities of channel availability of 0.6 ($\beta = 0.6$ and $\alpha = 0.4$) and 0.4 ($\beta = 0.4$ and $\alpha = 0.6$) and different values of M. The theoretical lower bound for $\mathrm{MTTF^h}$ for some cases is also included for comparison purposes. The results confirm the analysis in Section 4.5.1, which indicated that $\mathrm{MTTR^h}$ will exponentially converge toward $T + T_r$ with N and that $\log \mathrm{MTTF^h}$ linearly increases with N. It should also be noted that although the required CR capabilities and the channel access protocol are relatively simple, the number of channels required to achieve a high $\mathrm{MTTF^h}$ is relatively low compared to N_a. The results also clearly demonstrate the two reliability features (i.e., increasing the hard failures MTTF and decreasing the MTTR) that can be provided in an OSA wireless network using CRs with sensing and frequency switching capabilities in combination with a simple channel access protocol. Therefore, although increasing the number of channels is not spectrally efficient, it allows the CRN designer to arbitrarily decrease the occurrence of disconnection events and to decrease the disconnection time when such an event occurs. It thus provides more reliable service in unreliable channels, which is of paramount importance when providing communication infrastructure to support critical applications.

Furthermore, the results presented in Section 4.5 provide guidelines for the design of reliable wireless networks using OSA based on CRs. As can be seen in Fig. 4.9, there is an inflection point where $\mathrm{MTTR^h}$ stops decreasing significantly. Before this point, we can use the approximations for a loaded network derived in Section 4.5.2. From Eq. (4.14), we know that the best $\mathrm{MTTR^h}$ that can be achieved is $T + T_r$. Substituting this value into Eq. (4.26), we can find the following approximate threshold for the number of channels where the inflection point occurs :

$$N^*_{\mathrm{MTTR^h}} = \left\lceil \frac{M(\alpha+\beta)}{\beta} \right\rceil . \tag{4.27}$$

That is, we can use $N^*_{\mathrm{MTTR^h}}$ as the threshold between the loaded and lightly loaded network special cases. For example, for $M = 25$ and $\{\alpha, \beta\} = \{0.4, 0.6\}$ we obtain $N^*_{\mathrm{MTTR^h}} = 42$, and for $\{\alpha, \beta\} = \{0.6, 0.4\}$ we obtain $N^*_{\mathrm{MTTR^h}} = 63$. Substituting in Fig. 4.9, we can observe that for both cases we have $\mathrm{MTTR^h} < 3.2$, which confirms our analysis. Therefore, increasing N above $N^*_{\mathrm{MTTR^h}}$ will not significantly improve $\mathrm{MTTR^h}$. However, the value of $\mathrm{MTTF^h}$ in Fig. 4.10 is less than 100 for both cases ; this can be unacceptable in some situations. We can then use the linear approximation for the lightly loaded network for $\mathrm{MTTF^h}$ in Eq. (4.23) to

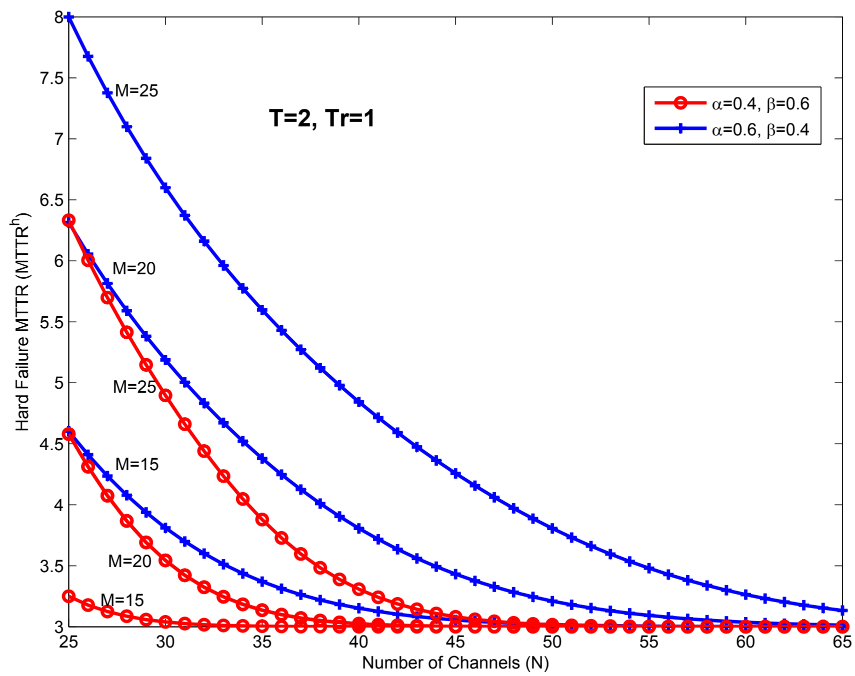

Figure 4.9 $\mathrm{MTTR}^{\mathrm{h}}$ versus N for different values of α, β and M.

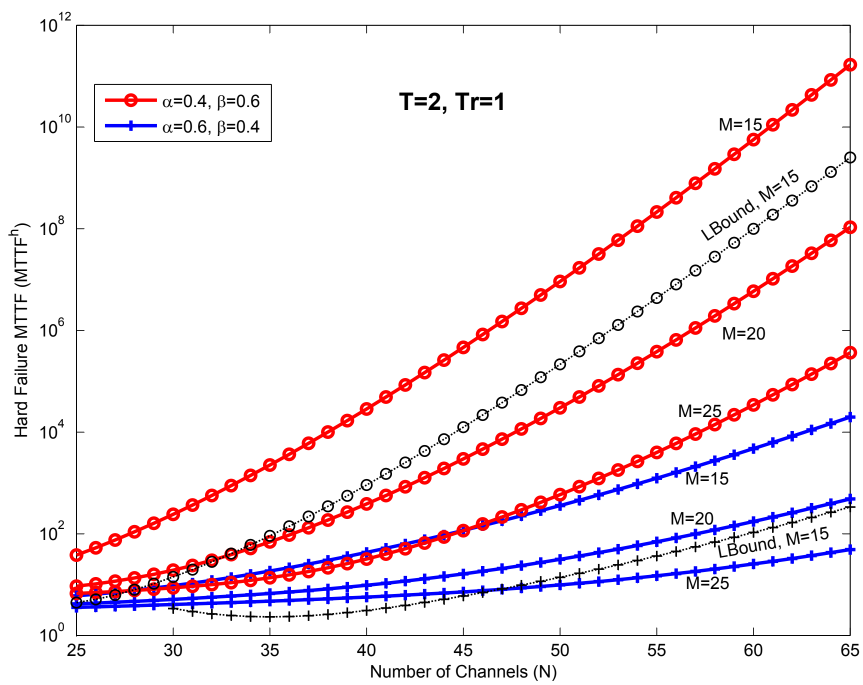

Figure 4.10 $\mathrm{MTTF}^{\mathrm{h}}$ versus N for different values of α, β and M.

compute an approximation for the number of channels required to achieve a given $\mathrm{MTTF}^{\mathrm{h}}$:

$$N^*_{\mathrm{MTTF}^{\mathrm{h}}} = N^*_{\mathrm{MTTR}^{\mathrm{h}}} + \left\lceil \frac{2\alpha \ln \mathrm{MTTF}^{\mathrm{h}}}{\beta} \right\rceil . \tag{4.28}$$

For example, for $M = 20$ and $\{\alpha, \beta\} = \{0.4, 0.6\}$ we obtain $N^*_{\mathrm{MTTR}^{\mathrm{h}}} = 53$ for a target $\mathrm{MTTF}^{\mathrm{h}} = 10^6$. This slightly underestimates the correct value of $N = 57$ because at $N^*_{\mathrm{MTTR}^{\mathrm{h}}}$ the linear approximation is not yet completely valid, and the slope at this point is less than the value used in Eq. (4.28). However, these design guidelines provide first order estimates and could also be used in optimization tools.

Although the availability can be as high as required when compared to $A_{\max}$ (Fig. 4.11), the availability is ultimately limited by $A_{\max}$ because it is not a function of N. This is due to the fact that soft failures still occur often even if the occurrence of hard failures is low because they are a function of the channel failures and because the MTTR for soft failure is always T_r. Therefore, the availability is limited by $(T_r)/T$. It is possible to increase T ; however, a longer T will increase $\mathrm{MTTR}^{\mathrm{h}}$, and the hypothesis that the channel does not change during the slot interval will no longer be valid.

4.6.1 Restoration Time

To make the analytical results tractable, we assumed a constant restoration (recovery) time equal to T_r. This assumption is accurate for the case where the channel assignment to the CR users is centralized. Moreover, the constant restoration time assumption is realistic when $N >> M$ because in this case the probability that multiple users will contend for the same channel is low. Therefore, T_r approaches a geometric distribution whose average will be :

$$\overline{T}_r = \left(1 + \frac{\alpha}{\beta}\right) \tau \tag{4.29}$$

where τ is the time spent to verify each channel and is approximately equal to the sensing time, which is an architectural parameter of the radio and is constant.

However, competition for access to available channels usually results in different restoration times for users. The *FAIL-THEN-CONTINUE* scheme Fan et Jiang (2009) illustrated in Figure 4.12 can be used to resolve the competition. In this scheme, access to the control channel to confirm the sensed available channel is divided into very short *microslots* that are used to resolve the competition using CSMA algorithms. These microslots are on the order of microsecond Fan et Jiang (2009) and the sensing time is on the order of milliseconds, so the impact of the collision for access to the same channel on the length of recovery minislots can be assumed to be constant. Therefore, the length of a minislot during the recovery, τ, is

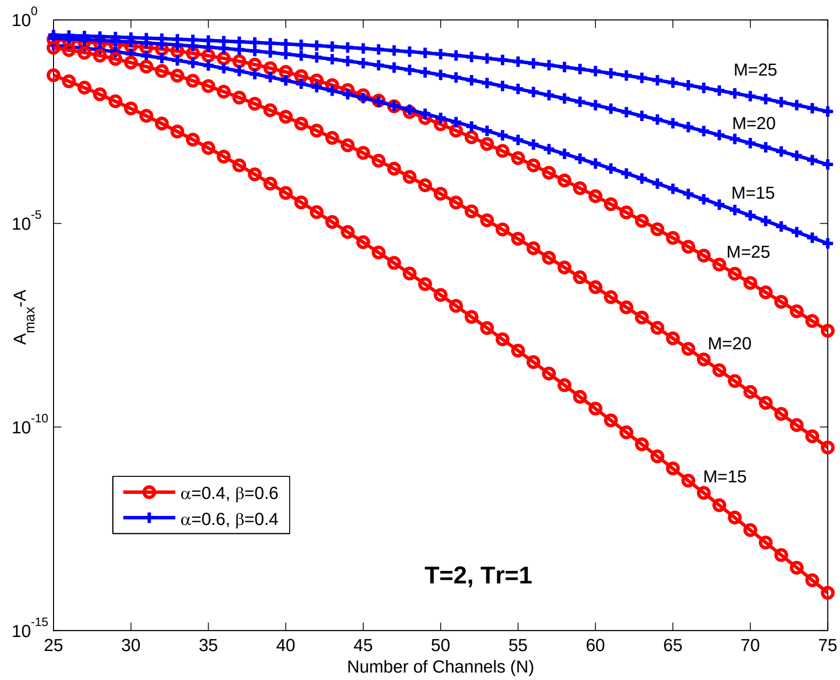

Figure 4.11 $A_{\max} - A$ versus N for different values of α, β and M.

composed of the sensing time, T_s, and a constant period for resolution of the competition, as can be seen at the bottom of Figure 4.12. For the example illustrated in Figure 4.12, three users follow the random sensing orders (321,123,321). Channel 1 is not available, but the two other channels are available. In the first minislot of the competition, it is assumed that the third user has won the competition, so its restoration time is equal to $T_s + \tau$. In the second minislot, the first user wins, and thus, its restoration time is equal to $T_s + 2\tau$. The second user is blocked in this timeslot. Therefore, the average restoration time can be given by $T_r = T_s + 1.5\tau$.

We simulated the CR network with random restoration times from the *FAIL-THEN-CONTINUE* contention resolution protocol and compared the simulation results to the analytical results obtained with a constant restoration time, whose value is given by the average restoration time of the *FAIL-THEN-CONTINUE* contention resolution protocol over all timeslots. The $\mathrm{MTTR}^{\mathrm{h}}$ and $\mathrm{MTTF}^{\mathrm{h}}$ reliability parameters obtained with the two approaches are compared in Figures 4.13 and 4.14. The results are similar even when $M = N$, which validates the constant restoration time simplification. Due to space limitations, we have only presented results for hard failures ; however, similar agreements were observed for all reliability parameters.

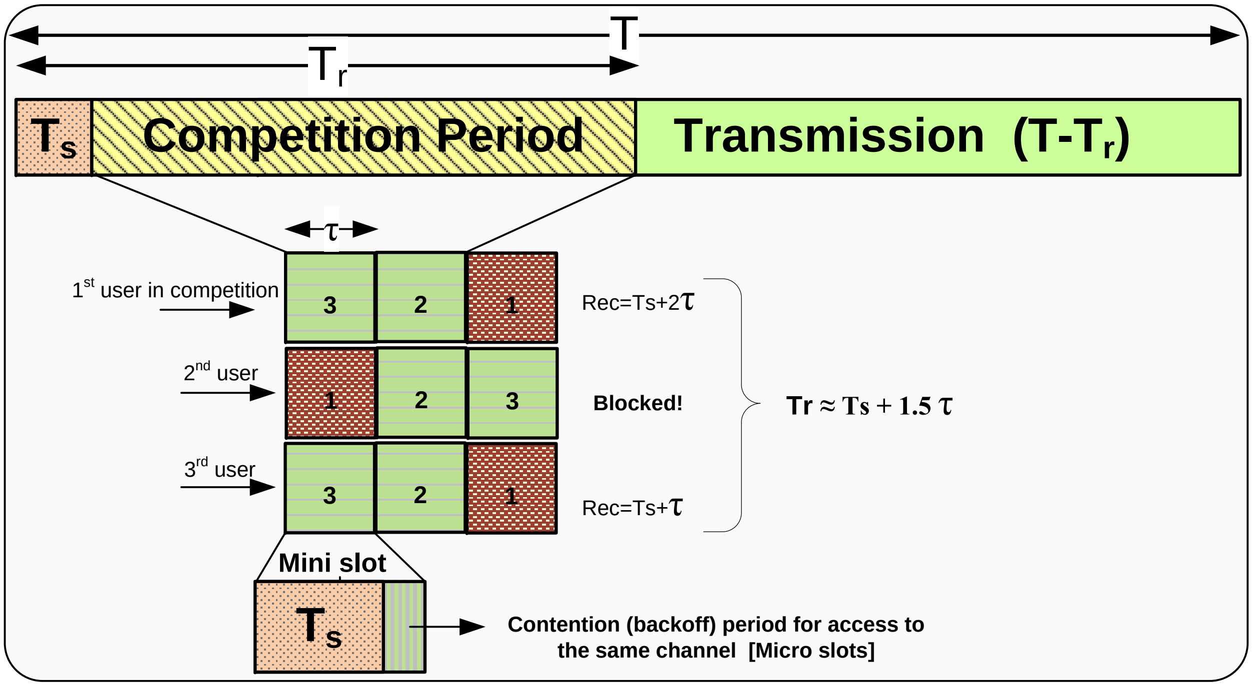

Figure 4.12 Interpretation of the constant restoration time for a random recovery model.

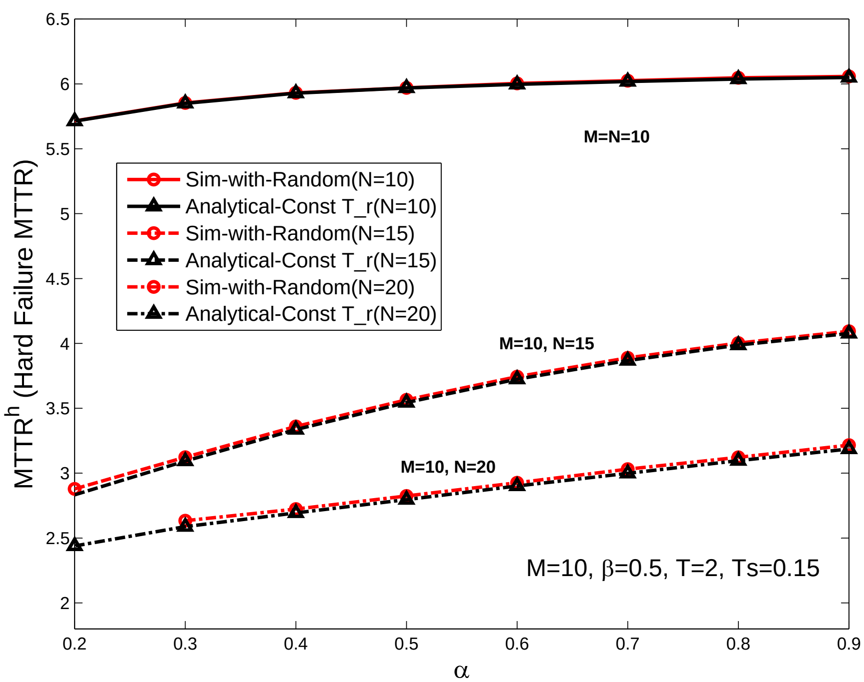

Figure 4.13 MTTR^h as a function of α for the analytical model with constant T_r and for the simulation model with the *FAIL-THEN-CONTINUE* contention resolution protocol.

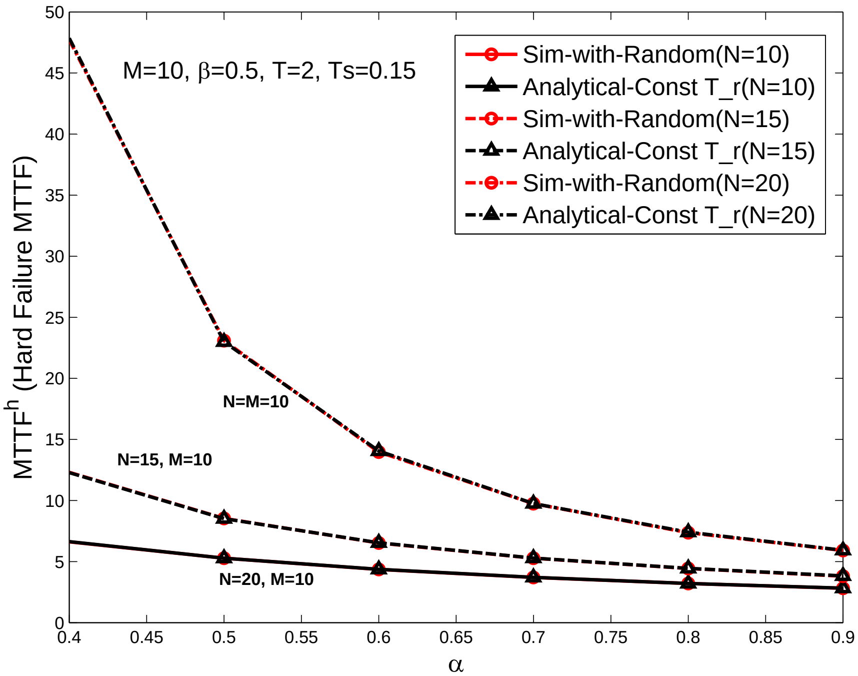

Figure 4.14 $\mathrm{MTTF}^{\mathrm{h}}$ as a function of α for the analytical model with constant T_r and for the simulation model with the *FAIL-THEN-CONTINUE* contention resolution protocol.

4.6.2 Impact of Sensing Errors

In this section, we discuss the impact of sensing errors on the reliability metrics. Sensing errors are modeled with two probabilities : the probability of miss-detection, P_m, which is the probability that an unavailable channel is sensed as available, and the probability of a false alarm, P_f, which is the probability of sensing an available channel as unavailable. The probability of detection, P_d, is $1 - P_m$.

Sensing errors were not addressed earlier in the analytical model because they are too complex to obtain analytical results. For example, in the presence of sensing errors, the sensing order of all competing users (see Figure 4.12) should be considered in order to compute the recovery time. The results in Fan et Jiang (2009) indicate that the analysis is complicated even for a two-user competition. We therefore limit our analysis to discussions and simulation results.

We first discuss the impact of in-band sensing errors for the users who held a channel in the previous slot and then discuss users involved in the recovery competition. When a user holds a channel, it performs in-band sensing at the beginning of the time slot. Two types of errors are possible. First, a miss-detection of channel unavailability may occur. A user in this situation will realize later that the channel is not available, but it is too late to participate in the recovery competition. It is therefore assumed that such users are blocked for the whole timeslot. In other words, a miss-detection triggers a hard failure, and the user is blocked. The user can also experience a false alarm that incorrectly triggers the recovery procedure. Furthermore, if the users share the sensing results of initial in-band sensing (as assumed in this paper), false-alarmed channels will be excluded from the competition. Thus, false alarms in the initial in-band sensing decrease the number of channels present in the competition and increase the number of competing users, which results in a higher blocking rate. False alarms thus trigger general failures that, based on the result of the competition, can be hard or soft failures.

Sensing errors result in a new three-state Markov chain for the users, as illustrated in Figure 4.15. Note that in this new model, the probability of success $P_{S|12}$ is also a function of the sensing errors, which are complex to model analytically. Moreover, the probability of success is different when the channel is not available and is correctly detected $(\alpha(1-P_m))$ and there is a false alarm $((1-\alpha)P_f)$ because the number of channels available in the competition is different. For this reason, two different notations for the probability of success have been employed ($P^{<}_{S|12}$ and $P^{>}_{S|12}$), where $P^{<}_{S|12}$ is the probability of success in the competition when the channel is correctly sensed as unavailable, and $P^{>}_{S|12}$ is the probability of success in the competition when the user has made a false alarm.

During the competition, a miss-detection terminates the recovery for that user as it starts using a channel that is unavailable. Such a user is thus blocked in this timeslot. False alarms during the competition do not exclude the channel from the sensing list of other users because no cooperation is assumed among users during the competition. However, the user that experiences a false alarm continues the recovery longer than required, and another user in the same minislot or later may sense this channel available and start using it.

Sensing errors thus decrease the MTTF and increase the MTTR. To evaluate the impact of sensing errors on the reliability parameters, simulations were performed for a scenario with fixed parameters (except the sensing errors) listed in Table 7.1. Due to space limitations, only the figures related to MTTR$^{\mathrm{h}}$ and MTTF$^{\mathrm{h}}$ are provided. For those parameters, the hard failure reliability metrics are $MTTR^h$=2.5 and $MTTF^h$=181.3 when there are no sensing errors ($P_f = P_m = 0$).

Figures 4.16 and 4.17 show $MTTF^h$ and $MTTR^h$, respectively, as a function of the sensing error probabilities P_f or P_m (the horizontal axis may represent only P_f with P_m=0, only P_m with P_f=0 and $P_f = P_m$). $MTTF^h$ decreases significantly for sensing error probabilities as low as 0.01. Furthermore, as a missed detection directly triggers a hard failure, $MTTF^h$ is affected more by the probability of miss-detection than by the probability of false alarms because a recovery may occur before being blocked in this case. On the other hand, $MTTR^h$ is more robust to sensing errors, with a significant increase only occurring for sensing error

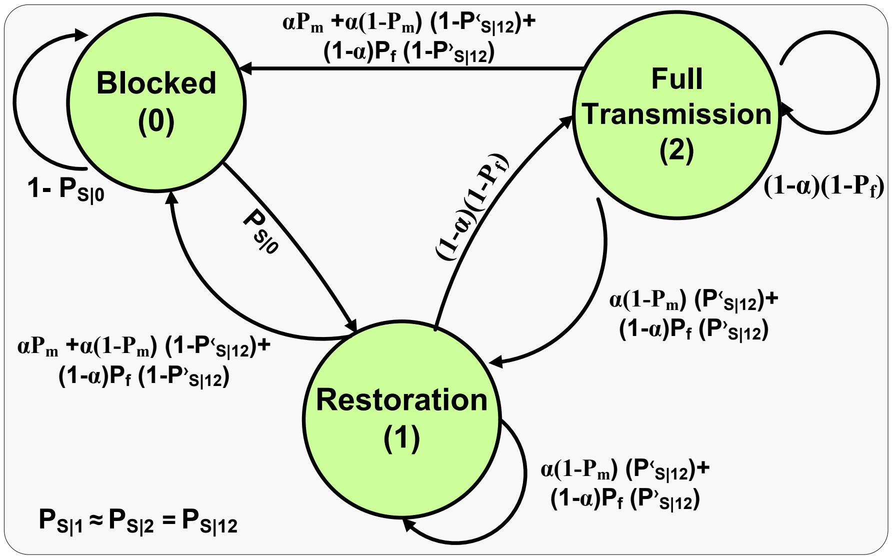

Figure 4.15 Three-state approximate Markov model for a CR user with sensing errors. The probabilities of success should also be found based on the sensing errors.

Table 4.1 Simulation parameters for the scenarios with sensing errors and variable restoration times for different users.

Notation	**Value**
T	2
T_s	0.15
N	25
M	15
β	0.7
α	0.3

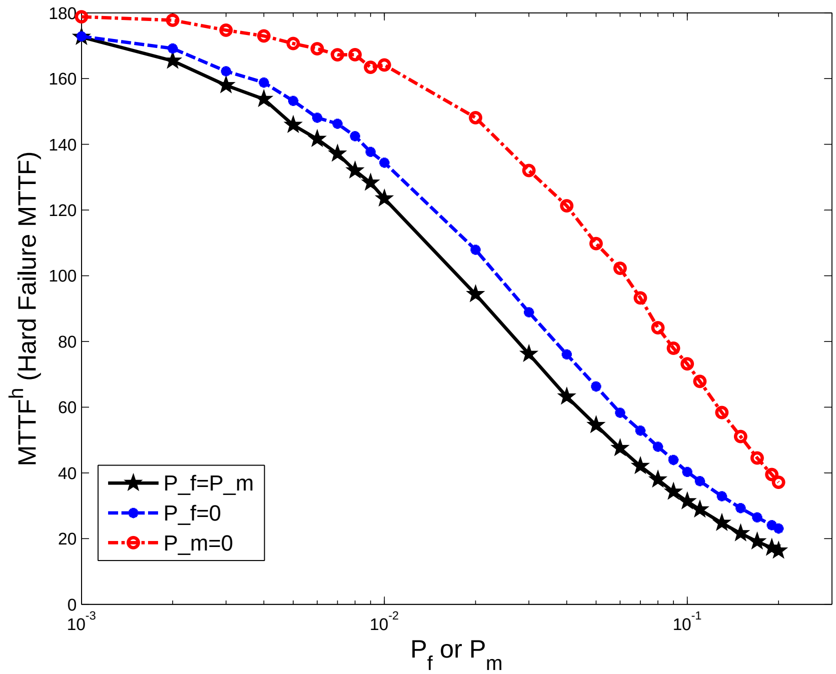

Figure 4.16 The impact of sensing errors on $MTTF^h$.

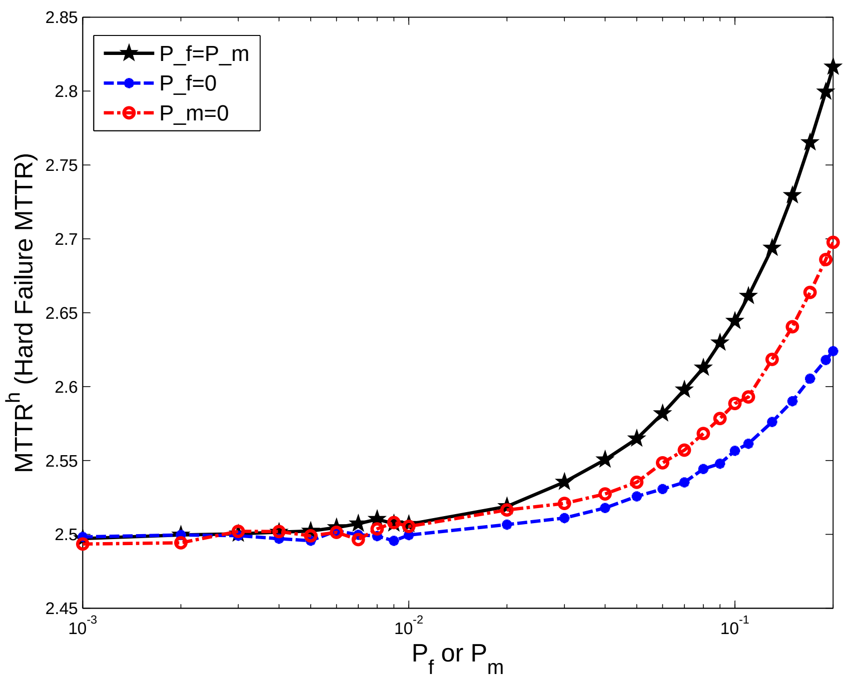

Figure 4.17 The impact of sensing errors on $MTTR^h$.

probabilities greater than 0.1. In conclusion, sensing errors will trigger hard failures more often than scenarios with perfect sensing, but the recovery time in presence of sensing errors is not significantly higher.

4.7 Conclusion

The design of reliable wireless networks, which is needed for critical applications, has not been thoroughly studied in the context of cognitive radios. In this paper, we presented a classification of several reliability parameters for wireless networks, such as MTTF, MTTR and availability, and provided closed-form relations for those metrics in a simplified network model. Using analyses of special cases, we then found simple expressions that relate the design parameters to the reliability metrics, which can be used as first-order design guidelines. While previous studies have mainly focused on the network level survivability, the results of this paper can be employed at the link level of such networks in conjunction with network level metrics, such as connectivity, to provide a comprehensive reliability model for cognitive radio networks. Finally, the results presented in this paper clearly demonstrate the capabilities of cognitive radios to provide reliable wireless network access and should serve as a motivation to pursue future research in this area, which includes the evaluation of reliability with unsaturated traffic, correlation between channels, fading channels, other channel occupancy distributions, the design of better channel access methods and the use of more advanced protection and restoration mechanisms. Further, the reliability analysis can also be extended for other classifications of failures, for instance, transient/permanent failures and single link/multi link failures Azarfar *et al.* (2012c).

Acknowledgments

This research project was supported by the Natural Sciences and Engineering Research Council of Canada under Strategic Grant STPG365205.

CHAPTER 5

ARTICLE 3 : PRIORITY QUEUEING MODELS FOR COGNITIVE RADIO NETWORKS WITH TRAFFIC DIFFERENTIATION

Arash Azarfar, Jean-François Frigon and Brunilde Sansò
Submitted to EURASIP Journal on Wireless Communications and Networking, April 2014
Department of Electrical Engineering
École Polytechnique de Montréal
C.P. 6079, succ. centre-ville, Montréal, QC, Canada, H3C 3A7
Email : {arash.azarfar, j-f.frigon, brunilde.sanso}@polymtl.ca

In this paper, we present a new queueing model providing the accurate average system time for packets transmitted over a cognitive radio (CR) link for multiple traffic classes with the preemptive and non-preemptive priority service disciplines. The analysis considers general packet service time, general distributions for the channel availability periods and service interruption periods, and a service-resume transmission. We further introduce and analyze two novel priority service disciplines for opportunistic spectrum access (OSA) networks which take advantage of interruptions to preempt low priority traffic at a low cost. Analytical results, in addition to simulation results to validate their accuracy, are also provided and illustrate the impact of different OSA network parameters on the average system time. We particularly show that, for the same average CR transmission link availability, the packet system time significantly increases in a semi-static network with long operating and interruption periods compared to an OSA network with fast alternating operating and interruption periods. We also present results indicating that, due to the presence of interruptions, priority queueing service disciplines provide a greater differentiated service in OSA networks than in traditional networks. The analytical tools presented in this paper are general and can be used to analyze the traffic metrics of most OSA networks carrying multiple classes of traffic with priority queueing service differentiation.

5.1 Introduction

Opportunistic spectrum access (OSA) is considered an important technology to address current and predicted exponential traffic growth in wireless networks Jondral (2007); Mitola

AZARFAR, A., FRIGON, J.-F. and SANSO, B. (2014) Priority queueing models for cognitive radio networks with traffic differentiation. Submitted to EURASIP Journal on Wireless Communications and Networking.

(2009); Xiao *et al.* (2013); Cisco (2013). Such growth is predominantly driven by multimedia traffic, such as video streaming Cisco (2013). Thus, it is expected that OSA networks will carry several traffic classes with different quality of service (QoS) requirements and importance.

The research objective of this paper is to obtain analytical tools to analyze traffic metrics, such as the packet system time, for differentiated services in opportunistic spectrum access networks. Such tools are required to evaluate the packet-level impact of OSA network parameters, novel medium access control (MAC) algorithms, channel sensing order strategies, etc. Moreover, those analytical tools can be used as a decision-making process for multimedia MAC algorithms Azarfar *et al.* (2012b), for OSA networks employing cognitive radio (CR) nodes, which possess learning and decision-making capabilities.

5.1.1 Related Work

Queueing models are the preferred approach to derive analytical results to analyze traffic metrics, and in Wang *et al.* (2011), the authors argue that queueing models can be the best choice to analyze a cognitive radio network. Priority service disciplines, such as the preemptive and non-preemptive service disciplines, are the most common approaches to implement service differentiation in communication networks. Furthermore, in an OSA network, the CR users must stop transmitting on an operating channel if the channel's primary user (PU) is detected or if the channel quality is unacceptable due, for example, to deep fading or interference. In the queueing model, the operating channel is the server of the queue. To achieve our objective, we must therefore analyze queueing models with priority service disciplines in the presence of frequent queue server interruptions.

Queueing models with preemptive priority service discipline and interruptions have been previously studied Federgruen et Green (1986); Avi-Itzhak (1963); Takagi (1991); D. P. Gaver (1962); Fiems *et al.* (2008). Some of that work considered that interruption periods are server busy periods generated by higher priority classes of traffic. For this approach, the interruption periods are not generally distributed since they depend on the arrival rate and service rate of higher priority classes. Obtaining the interruption period distribution is therefore not always straightforward. Inversely, given an interruption period distribution, it is not easy to find the appropriate arrival and service processes whose busy period has this distribution. The other articles that studied generally distributed interruption periods only considered a single traffic class and only provided bounds.

Few have attempted to provide queueing models for opportunistic spectrum access networks. In Wang *et al.* (2011); Rashid *et al.* (2007); Laourine *et al.* (2010); Wang *et al.* (2010), queueing models for an OSA network with a single class of traffic were derived using a similar approach as in Federgruen et Green (1986); Avi-Itzhak (1963); Takagi (1991); D. P. Gaver

(1962); Fiems *et al.* (2008) whereby the server interruption periods for the cognitive radio users are busy periods generated by the preemptive primary traffic. This approach has several major deficiencies. First, those models are limited to exponential operating period length. Also, as previously discussed, they can not address arbitrary interruption period lengths before the transmission can resume. This is particularly important for OSA networks since the interruption period length depends on several factors such as the MAC policy, the number of available channels, the number of competing CR users, etc. Even for the simple case where CR users wait until the PU releases the channel, the interruption period is not necessarily distributed as the busy period of a PU Poisson traffic Chen *et al.* (2009). The approach of simply considering the interruption periods as a preemption from PU traffic is therefore not accurate and general enough to analyze OSA networks.

In Azarfar *et al.* (2012a) we addressed several of those problems in a new queueing model for a single class of CR traffic for general operating and interruption period lengths. However, this model was limited to constant service time. In Li et Han (2011), an optimal threshold for the queue length to decide whether a packet should join the queue or not is derived. However, the model is again not general and can not be used to analyze traffic metrics.

To the best of our knowledge Kim (2012) is one of the few papers discussing a queueing system with multiple classes of traffic in cognitive radio networks. The authors analyze a T-preemptive scheme and, similarly to the other work on opportunistic spectrum access networks, the queueing analysis does not consider general interruption lengths and it is specific to the priority service disciplines considered. In Shiang et van der Schaar (2008), the authors consider a queueing model with multiple classes and a single priority service discipline for cognitive radios to address the problem of channel selection.

5.1.2 Contributions

In this paper, we consider real server interruptions distinct from the service times for a high priority class as well as general service time. We thus present, to the best of our knowledge, the first queueing model providing the accurate average system time for a Poisson packet arrival process with general service time transmitted over a CR link with general interruption periods and exponentially distributed operating periods for both a single traffic class and for multiple traffic classes with the preemptive and non-preemptive priority service disciplines. We also derive an approximate analysis for general operating period distributions. We further introduce two novel priority service disciplines which are specific to OSA networks with service interruptions. In the first novel OSA service discipline that we name *exceptional non-preemptive*, the service is in general non-preemptive except for low priority arrivals in an empty queue during an interruption period, which can be preempted by high priority packet arrivals during the same interruption period. In the second novel OSA service discipline that we name *preemptive in case of failure*, the service is non-preemptive during the operating periods but high priority packets can preempt low priority packets at the end of an interruption period. We provide an accurate analysis for the first novel OSA priority service discipline while approximate results are provided for the second (exact results are derived for the preemptive in case of failure service discipline for exponential service times). Since no specific assumptions are made regarding the nature of the operating and interruption periods, the results and derivations presented in this paper can be used to analyze the traffic metrics of most OSA networks with different MAC protocols. The final contributions of this paper are new insights on OSA networks based on the average system time analysis. Particularly, we show that, for the same average CR transmission link availability (ratio between average channel availability period length and average interruption period), the packet system time significantly increases as the operating and interruption periods average length exceeds the packet service time. We also present results emphasizing the critical importance of minimizing the interruption period lengths to minimize the packet system time in OSA networks. Another conclusion that we present is that priority queueing service disciplines provide a greater differentiated service in OSA networks than in traditional networks.

The reminder of the paper is organized as follows. In Section 5.2 we present the cognitive radio system and the queue model. In Section 5.3, an M/G/1/ queue with interruptions and with a single class of CR traffic is discussed. The results are then used in Section 5.4 to analytically solve four priority queueing disciplines in the presence of interruptions. We also present in Section 5.5 an alternative approach to analyze the preemptive and non-preemptive disciplines for exponential operating periods. Analytical and simulation results are presented in Section 5.6 and finally Section 5.7 concludes the paper with some remarks on future research directions.

5.2 Cognitive Radio Queue Model

The cognitive radio queue model can be summarized as follows. We consider a pair of cognitive radio users operating using opportunistic spectrum access over one or more wireless channels. During an operating period, packets that are in one of the CR nodes queue are transmitted to the other CR node according to a chosen service discipline. As illustrated in Fig. 5.1, the CR nodes opportunistically operate over a channel for a random duration Y until the channel becomes unavailable. When the channel becomes unavailable, the packet transmission is interrupted for a random length R until an available operating channel can be used by the CR pair, at which time the packet transmission is resumed.

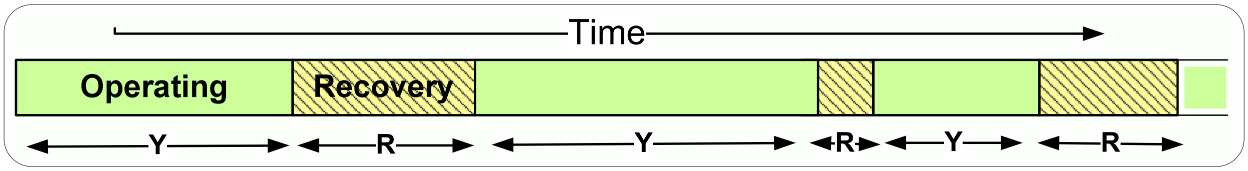

Figure 5.1 Operation model for a cognitive radio link alternating between operating and recovery (interruption) periods. Identical instances of Y and R are illustrated.

We now describe the details of the model. The OSA network assigns a channel to the pair of CR users according to its MAC protocol and channel assignment algorithm. Transmission over the assigned channel can be multiplexed with other CR users, the only assumption for our model is that during channel availability periods, the pair of CR users have access to a constant service rate over the assigned channel (the packet length defines the distribution of the real service time). If, as in IEEE 802.22, periodic quiet periods are required to sense the channel or perform other OSA network tasks, the channel service rate can be scaled accordingly. Channels are assumed to be homogeneous with the same service rate. The CR nodes communicate over the assigned channel for a random operating duration Y until the channel becomes unavailable and packet transmissions must be stopped. We denote the instant where the channel becomes unavailable for operation as a failure event Azarfar *et al.* (2012c). To illustrate the generality of this model, we now give a few examples of failure events. A failure can be due to the appearance of the primary user, a false detection of the primary user, a link failure due to excessive transmission loss (fading, shadowing or distance), or interference from other secondary users. A failure event can also be due to the OSA protocol. For example, CR users might have to release a channel after a fixed period of time, even if no primary user appears. Note that it is also implicitly assumed that no miss-detection may occur, so there will be no performance degradation for primary users to be analyzed.

For the model and its analysis, only the distribution of Y is required and the exact reason for the failure event is irrelevant as long as it is independent of the packet transmission process (e.g., the pair of CR users are not reassigned to a new channel after each packet transmission or when the packet queue is empty). Note that when the CR users start using a channel, unless it is immediately after a channel unavailability period, they generally have no knowledge about how long this channel has been available. Therefore, Y is a function of the residual time of the availability period of the channel Azarfar *et al.* (2012a).

The recovery or interruption period denotes the period of time R during which the CR users can not transmit and try to recover the transmission Azarfar *et al.* (2012c). The length of R depends on the OSA network model but only its distribution is relevant for the queue analysis. We will use a few examples of recovery periods to demonstrate the generality of the proposed model. For OSA MAC protocols in which the CR users buffer the packets until the operating channel becomes available again Park *et al.* (2011), the distribution of R is identical to the channel unavailability period distribution. For network with a channel switching policy in which when the channel becomes unavailable, the CR users enter a competition with other CR users to be granted access to a new channel Park *et al.* (2011), the distribution of R will depend on the MAC competition protocol (e.g., slotted Aloha), the number of users, the number of available channels, etc. Even if the user is blocked due to other users transmitting on all channels, the total time of blocking until a successful channel reservation is included in the recovery time. For OSA networks where a channel is granted by a spectrum server, the length of R can be a fixed period of time (query and service time, radio switching time, etc.). A methodology to find the recovery period distribution for two baseline multichannel opportunistic spectrum access MAC protocols is provided in Azarfar *et al.* (2014a).

To summarize, determining the distribution of Y and R according to the OSA network model under study is outside the scope of this paper. But once the distributions are known, the queue model that we are presenting can be used to find the traffic metrics for the OSA network CR users.

As illustrated in Fig. 5.2, we consider a CR system with N traffic classes where each class i, $i = 1, \ldots, N$, has an independent Poisson packet arrival process with rate λ_i and the total arrival rate is $\lambda = \sum \lambda_i$. We also denote by A_i, the inter-arrival time between packets of class i, $i = 1, \ldots, N$, and define $\mathbb{A} = \min\{A_1, ..., A_N\}$ as the inter-arrival time between packets in the system. Throughout the paper, for any random variable Z, $f_Z(.)$ and $F_Z(.)$ respectively represent the probability density/mass function (PDF or PMF) and the cumulative distribution function (CDF) of the random variable Z. Moreover, $\widehat{Z}(s)$ represents the Laplaceñ-Stieltjes transform (LST) of the distribution $F_Z(.)$ of the random variable Z.

Lower index classes have higher priority. In the special case of two traffic classes (e.g., voice and data), we designate the index 1 traffic as high priority (HP) and the index 2 traffic as low priority (LP). Packets from traffic class i have a random *real* service time T_i. The real

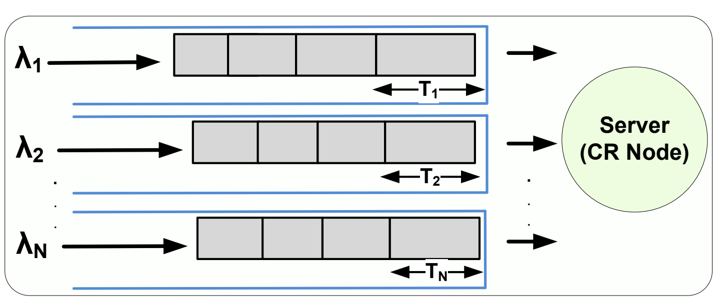

Figure 5.2 Queue model with a multiple-class cognitive radio traffic.

service time is the total transmission time of the packet and excludes the time spent during interruption periods during the service of a packet. From the queueing point of view, the user's operation can thus be modeled as an M/G/1 queue with random service interruptions.

We must also introduce the notion of *completion time* X, which represents the whole time in service for a packet including the real service time T and the interruptions that may occur during its service. We assume a service-resume model which means that after a packet service interruption, only the remaining part of the packet needs to be transmitted. This implies that the completion time of a packet is formed by alternating instances of Y and R named $Y_1, Y_2, \ldots$ and $R_1, R_2, \ldots$ respectively. The queue size is assumed infinite, so packet loss and blocking are irrelevant and the main performance metrics are the total time spent in the queue (waiting time) W and in the system (system time or sojourn time) $D = X + W$.

We consider four different service disciplines : the classical *non-preemptive* and *preemptive-resume* schemesBertsekas et Gallager (1992), and two novel disciplines we propose in this paper. As illustrated in Fig. 5.3, if during a recovery period a low priority (LP) packet arrives in an empty system followed by a high priority (HP) packet, in a non-preemptive scheme the LP packet will be transmitted first. In other words, the LP packet can not be preempted even if its real service has not started yet. In the new scheme that we call *exceptional non-preemptive*, an HP packet can preempt a lower priority packet only if its real service has not started yet. As can be seen in Fig. 5.3, the difference between a non-preemptive and an exceptional non-preemptive scheme is only for the LP packets which arrive to an empty unavailable system. We also propose a *preemption in case of failure* discipline where HP packets can not preempt an LP packet in service until the LP service is finished or if an interruption occurs. In other words, at the end of a recovery period, the priority is always given to HP packets. Meanwhile, in the classical preemptive scheme HP packets can preempt LP packets at any time. The two proposed schemes are defined based on the existence of interruptions and can specifically be used in OSA networks with service interruptions.

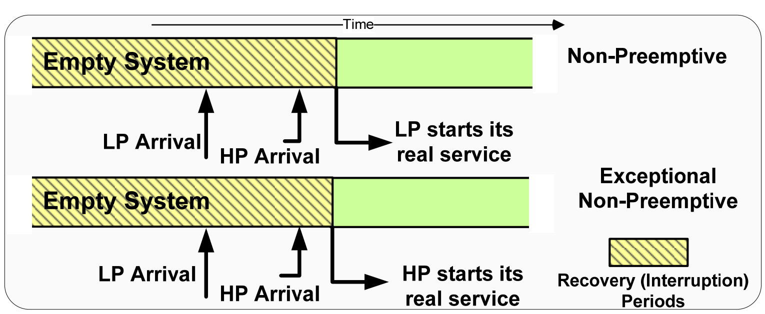

Figure 5.3 Comparison between non-preemptive and exceptional non-preemptive schemes.

5.3 Single Traffic Class Analysis

In this section, we analyze the queue model with a single traffic class and obtain results which will be extended in the next section to the analysis of multiple classes of traffic. The results presented here are an extension of the work presented in Azarfar *et al.* (2012a) where only constant service time was considered. In this section, the analysis is done for the general case where the packet length follows an arbitrary random distribution.

As can be seen in Fig.'s 5.4 and 5.5, we can distinguish three types of packet in the CR queueing model. There are packets which enter an empty available system and their real service starts immediately (see Fig. 5.4.1). The subscript 'a' is used to designate this case. Considering the completion time of the packets of type 'a', we can see that the distribution of the first operating period is different from the following ones because it represents the residual part of Y. We thus use the notation of Y_{1a} to designate the first operating period for the case 'a' packets.

Packets that enter an empty system during an interruption period (empty unavailable system) must wait until the end of the recovery period before starting their real service (see Fig. 5.4.2). The subscript 'u' is used to designate this case. For those packets, the distribution of the first operating period is the same as the operating period distribution and is simply denoted by Y_1. On the other hand, the distribution of the first recovery period is different from the following recovery periods because it represents the residual part of R. R_r is used to denote the remaining part of the recovery period in which the arrival has occurred. Note that we consider the arrival time as the start of the completion time in case 'u' (i.e., R_r is not accounted as waiting time but as completion time).

Finally, there are packets that enter a busy system and are queued (see Fig. 5.5). Their service starts immediately after the completion time of the previous packets and the subscript 'b' is used to designate this case. Note that in this case, the completion time of the packet

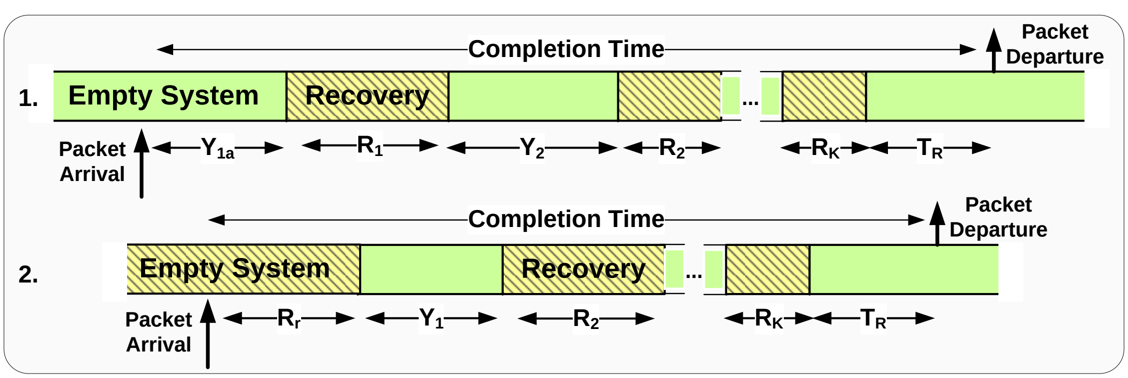

Figure 5.4 Completion time for the case 'a' (1.) and case 'u' (2.).

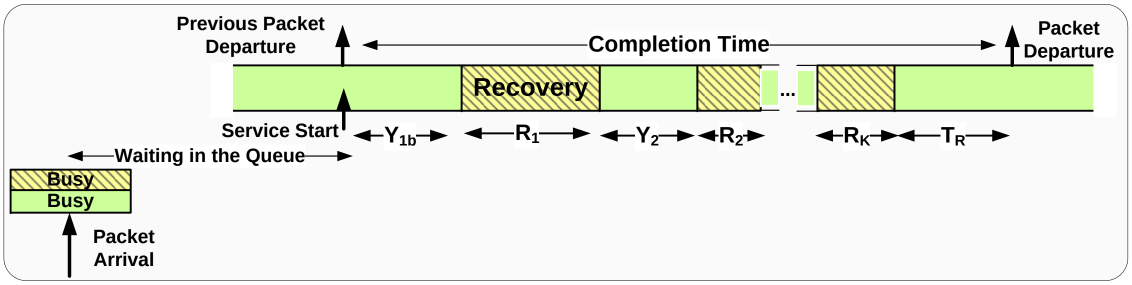

Figure 5.5 Completion time for the third case when the packet enters a busy system and is queued (case 'b').

is always started within an operating period (similar to the case 'a'). Their first operating period is thus called Y_{1b} because its distribution is different from general Y.

In the following, we will find the first two moments of the completion time X and then analyze the average waiting time and other metrics. We also provide simplifications for the special case of exponential operating periods.

5.3.1 Completion time

Suppose X_a represents the completion time of the packets of type 'a' (similarly X_b and X_u for cases 'b' and 'u'). Let also X_e define the completion time of a packet which arrives to an empty system (cases 'a' and 'u' together). The first two moments of X_e are given by :

$$E[X_e] = P_{ae}E[X_a] + (1 - P_{ae})E[X_u], \tag{5.1}$$

$$E[X_e^2] = P_{ae}E[X_a^2] + (1 - P_{ae})E[X_u^2], \tag{5.2}$$

where P_{ae} is the average probability that the server is available when the system is empty. From Federgruen et Green (1986), P_{ae} can be found equal to :

$$P_{ae} = 1 - \frac{(1 - \widetilde{F}_Y(\lambda))(1 - \widetilde{F}_R(\lambda))}{\lambda E[Y](1 - \widetilde{F}_Y(\lambda)\widetilde{F}_R(\lambda))}, \tag{5.3}$$

where $\widetilde{F}_Z(.)$, for an arbitrary distribution function $F_Z(.)$, is given by :

$$\widetilde{F}_Z(\lambda) = \int_0^\infty e^{-\lambda t} dF_Z(t). \tag{5.4}$$

For $Z = Y$ or R, this function gives the probability that the length of the operating or recovery period, respectively, is less than a packet inter-arrival time. Note that P_{ae} is a conditional probability, conditioned on the fact that the system is empty. In general, P_{ae} is only a function of the moments of Y and R. We therefore assume in the following that there is no correlation between P_{ae} and Y and R.

The first moment of X can be obtained by solving the following equation :

$$E[X] = \rho E[X_b] + (1-\rho)E[X_e], \tag{5.5}$$

where $\rho = \lambda E[X]$ is the probability of the system being not empty. The second moment of X can be found equal to :

$$E[X^2] = \rho E[X_b^2] + (1-\rho)E[X_e^2]. \tag{5.6}$$

We will now find the first two moments of X_a, X_u and X_b.

Arrival to an empty-available system

For packets of case 'a', as illustrated in Fig. 5.4.1, X_a, can be given by :

$$X_a = \begin{cases} T, & Y_{1a} \geq T \\ Y_{1a} + R_1 + Y_2 + R_2 + ... + Y_K + R_K + T_R, & \text{Otherwise}, \end{cases} \tag{5.7}$$

where Y_{1a} is the random remaining time of the first operating period until the next interruption, T_R is the transmission time of the last part of the packet, K is the number of operating periods required to transmit the entire packet and

$$T = Y_{1a} + ... + Y_K + T_R. \tag{5.8}$$

If we consider the operating periods $\{Y_{1a}, Y_2, \ldots, Y_K\}$ as a renewal process, K is the number of renewals of Y during the real transmission time of a packet and its distribution can be found from the renewal theory results Cox (1962); Ross (2006). The first two moments of the number of renewals during $(0, t]$ composed of instances of Y_{1a} and Y are given by :

$$m_a(t) = \mathcal{L}^{-1}\left\{\frac{\widehat{Y}_{1a}(s)}{s(1-\widehat{Y}(s))}\right\}, \tag{5.9}$$

$$m_a^2(t) = \mathcal{L}^{-1}\left\{\frac{\widehat{Y}_{1a}(s)(1+\widehat{Y}(s))}{s\left(1-\widehat{Y}(s)\right)^2}\right\}. \tag{5.10}$$

The moments of K_a are then given by :

$$E[K_a] = \int_0^\infty E[K_a|T=t]f_T(t)dt = \int_0^\infty m_a(t)f_T(t)dt. \tag{5.11}$$

We can rewrite (5.7) as :

$$X_a = T + \sum_{k=1}^{K} R_k. \tag{5.12}$$

We then obtain the first moment of X_a as :

$$E[X_a] = E[T] + E[K_a]E[R]. \tag{5.13}$$

For the second moment, we use the fact that K_a is independent of the recovery process, but not the service time, and that the variance of the random sum $\sum_{k=1}^{K} R_k$ is equal to $E[K]Var(R) + (E[R])^2 Var(K)$ to obtain that

$$E[X_a^2] = E[T^2] + 2E[TK_a]E[R] + E[K_a](E[R^2] - (E[R])^2) + (E[R])^2 E[K_a^2], \tag{5.14}$$

where

$$E[TK_a] = \int_0^{\infty} E[TK_a|(T = t)]f_T(t)dt = \int_0^{\infty} tm_a(t)f_T(t)dt. \tag{5.15}$$

Arrival to a busy system

Based on the distribution of the first operating period Y_{1b} (see Fig. 5.5), we can find the first two moments of K_b, the number of renewals for the case 'b', and X_b , as we did for K_a and X_a in the previous results.

Arrival to an empty-unavailable system

For the case 'u', as illustrated in Fig. 5.4.2, the completion time of the user is started within a recovery period and we have :

$$X_u = R_r + Y_1 + R_2 + ... + Y_K + R_{K+1} + T_R = R_r + X_u^*. \tag{5.16}$$

The moments of X_u^* can be found as for X_a by replacing Y_{1a} with Y_1. R_r, the remaining time in the first recovery period, can be written as :

$$R_r = R - \mathbb{A}|(R > \mathbb{A}), \tag{5.17}$$

where $\mathbb{A}|(R > \mathbb{A})$ is the inter-arrival time conditioned on the fact that it should be less than R. The moments of R_r based on R and $\mathbb{A}$ are found as follows.

As above, we encounter several times throughout the paper the random variables $Z = V|(V < U)$ and $Q = V - U|(V > U)$, for any two arbitrary random variables V and U. We derive here the statistics of these two random variables. We have :

$$f_Z(t) = \frac{Pr(U > t) f_V(t)}{Pr(U > V)} = \frac{(1 - F_U(t)) f_V(t)}{Pr(U > V)}. \tag{5.18}$$

When U is exponentially distributed with parameter α, we then have :

$$Pr(V < U) = \int_0^\infty e^{-\alpha t} f_V(t) = \widehat{V}(\alpha), \tag{5.19}$$

$$f_Z(t) = \frac{e^{-\alpha t} f_V(t)}{\widehat{V}(\alpha)}. \tag{5.20}$$

In this case, $E[Z]$ and $E[Z^2]$ can respectively be given by :

$$E[Z] = \frac{-d/d\alpha \widehat{V}(\alpha)}{\widehat{V}(\alpha)}, \tag{5.21}$$

$$E[Z^2] = \frac{d2/d\alpha^2 \widehat{V}(\alpha)}{\widehat{V}(\alpha)}. \tag{5.22}$$

For the second random variable, Q, we still assume that U is exponentially distributed with parameter α. Then, after some algebra manipulations (details can be found in (Federgruen et Green, 1986, Lemma2) or in Takagi (1991)), we obtain :

$$E[Q] = \frac{E[V]}{1 - \widehat{V}(\alpha)} - \frac{1}{\alpha}, \tag{5.23}$$

and

$$E[Q^2] = \frac{E[V^2] - 2\frac{E[V]}{\alpha}}{1 - \widehat{V}(\alpha)} + \frac{2}{\alpha^2}. \tag{5.24}$$

Based on the moments of $R_r = R - \mathbb{A}|(R > \mathbb{A})$, the completion time can thus be found as :

$$\begin{aligned} E[X_u] &= E[R_r] + E[X_u^*], \\ E[X_u^2] &= E[R_r^2] + E[X_u^{*2}] + 2E[R_r]E[X_u^*]. \end{aligned} \tag{5.25}$$

5.3.2 Queue performance metrics

We can use the same approach used for M/G/1 queues to derive the waiting time for our system. When a packet arrives, it waits for the remaining completion time of the packet in service (if any), and then the completion time of all packets in the queue. For the packets which are in the queue, the completion time is always distributed with X_b (they are queued, so they have not arrived to an empty system). However for the packet which is initially in service, the general completion time should be used because no knowledge is available to know whether this packet has been of case 'a', 'b' or 'u'. We thus have :

$$E[W] = \frac{\lambda E[X^2]}{2(1 - \lambda E[X_b])}. \tag{5.26}$$

The average system time is given by $E[D] = E[W] + E[X]$.

5.3.3 Busy periods

Similarly to an M/G/1 queue without interruption Takagi (1991), we can find the busy periods' distribution for our queue with interruption. This result will be useful to analyze the priority disciplines. We know that the first completion time in a busy period is an instance of X_e. However, for other busy periods which are initiated during X_e, the busy period is started with an instance of X_b because the packets enter a non-empty system. Therefore, we can find the LST of the busy periods as :

$$\widehat{B}(s) = \widehat{X}_e(s + \lambda - \lambda \widehat{B}_b(s)), \tag{5.27}$$

where $\widehat{B}_b(s)$ is the LST of the busy periods which are initiated during X_e with an instance of X_b. $\widehat{B}_b(s)$ itself can be found from the following equation :

$$\widehat{B}_b(s) = \widehat{X}_b(s + \lambda - \lambda \widehat{B}_b(s)). \tag{5.28}$$

From the equation above, we can find the first and the second moments of $B_b(t)$ as :

$$E[B_b] = \frac{E[X_b]}{1 - \lambda E[X_b]} \text{ and } E[B_b^2] = \frac{E[X_b^2]}{(1 - \lambda E[X_b])^3}. \tag{5.29}$$

The first and the second moments of the general busy periods are then given by :

$$E[B] = \frac{E[X_e]}{1 - \lambda E[X_b]}, \tag{5.30}$$

$$E[B^2] = \lambda E[B_b^2] E[X_e] + (1 + \lambda E[B_b])^2 E[{X_e}^2]. \tag{5.31}$$

5.3.4 Alternative model

An alternative model is to consider the start of the real service as the start of the completion time. We introduce this model since it will be useful to analyze some of the priority schemes. This alternative model does not affect the completion time for arrivals in empty available and busy systems (cases 'a' and 'b'), but for an arrival in an empty unavailable system, the remaining time of the recovery period R_r is considered as waiting time. The completion time is given by X_u^* (see (5.25)). The first two moments of X^*, the overall completion time for this alternative model, can then be found using the same approach as for X. Since the system time for both models must be the same we then have that the average waiting time for this model is given by :

$$E[W^*] = E[W] + E[X] - E[X^*]. \tag{5.32}$$

5.3.5 Approximate and exponential operating periods

As discussed in Federgruen et Green (1986), in general it is very complex to find the exact distribution of Y_{1a} and Y_{1b}, since they depend on the time when a packet arrives or a packet service has terminated. An approximation for Y_{1a} and Y_{1b} is to assume that they may be started uniformly during an operating period Azarfar *et al.* (2012a), which is sometimes called *random modification* of Y Avi-Itzhak (1963) or *equilibrium excess distribution* Federgruen et Green (1986).

For the special, yet important, case that the operating periods are distributed with an exponential distribution Y with parameter α (i.e., $F_Y = 1 - e^{-\alpha t}$), from the memoryless property we have that $Y_{1a} = Y_{1b} = Y$ and $K_a = K_u = K_b$. Using (5.9) and (5.10), we have that $m(t) = \alpha t$ and $m^2(t) = \alpha^2 t^2 + \alpha t$. It is then straightforward to derive the moments of the completion time as :

$$E[X_{a,b,u}] = E[T](1 + \alpha E[R]), \tag{5.33}$$

$$E[TK_a] = \alpha E[T^2], \tag{5.34}$$

$$E[X_b^2] = E[T^2](1 + \alpha E[R])^2 + \alpha E[T]E[R^2], \tag{5.35}$$

$$E[X_u^2] = E[X_a^2] + E[R_r^2] + 2E[X_a]E[R_r]. \tag{5.36}$$

For exponentially distributed operating periods, we can further model our queue as a queue with an initial setup time Takagi (1991) to find the waiting time. The initial setup time S for a packet which initiates the busy period is R_r with probability $(1 - P_{ae})$ and zero otherwise. Thus, we can find the moments of S based on the moments of Y and R :

$E[S] = (1 - P_{ae})E[R_r]$ and $E[S^2] = (1 - P_{ae})E[R_r^2]$. From (Takagi, 1991, (2.44a)), we then have :

$$\begin{aligned} E[D] &= E[X_b] + \frac{\lambda E[X_b^2]}{2(1 - \lambda E[X_b])} + \frac{2E[S] + \lambda E[S^2]}{2(1 + \lambda E[S])} \\ &= E[X_b] + \frac{\lambda E[X_b^2]}{2(1 - \lambda E[X_b])} + \frac{E[R^2]}{2(E[Y] + E[R])}. \end{aligned} \tag{5.37}$$

The steady-state probability of the system being empty, P_0, can be given by :

$$P_0 = \frac{E[I]}{E[I] + E[B_s]} = \frac{1 - \lambda E[X_b]}{1 + \lambda E[S]}, \tag{5.38}$$

where $E[I] = \frac{1}{\lambda}$ is the average of idle periods (no packet in the system), and $E[B_s]$ is the average of busy periods initiated by $S + X_b$ which can be found from Section 5.3.3.

5.3.6 Case Study : Comparison Between Switching and Buffering OSA Strategies

We now present a case study to validate the theoretical analysis and to discuss how it can be used to gain insight on the performance of OSA networks. In this case study, we compare two common OSA strategies which, following the detection of primary users activity on the operating channel, either *switch* to a new channel or *buffer* packets while waiting for the primary users to release the channel Park *et al.* (2011); Lai *et al.* (2011).

It is assumed that there is a large set of similar channels with exponentially distributed availability (I) and unavailability periods (U). For both OSA policies, we have $Y = I$ and for the buffering policy $R = U$ Azarfar *et al.* (2012a). For the OSA switching policy, we use the common random sensing model in which the channels are sensed successively in a random order until an available channel is found. The interruption time R is thus geometrically distributed with a success probability $\frac{E[I]}{E[I]+E[U]}$ for each time slot of τ (τ is the amount of time required to switch to and sense a channel). In the theoretical model, R is approximated by an exponential distribution with an average length $E[R] = \frac{\tau(E[I]+E[U])}{E[I]}$.

Figure 5.6 compares the packet sojourn time ($E[D]$), for the system parameters indicated in the figure, of these two models obtained with exact Monte-Carlo simulations and using the theoretical result (5.37). First, the presented results confirm the accuracy of the theoretical model and its applicability to different OSA strategies. We can also observe that, as can be expected, the threshold point for the average channel unavailability length $E[U]$ where the switching policy becomes preferable over the buffering policy increases from 5.2 to 11.4 units of time when τ increases from 5 to 10 units of time. Note that this threshold is not simply

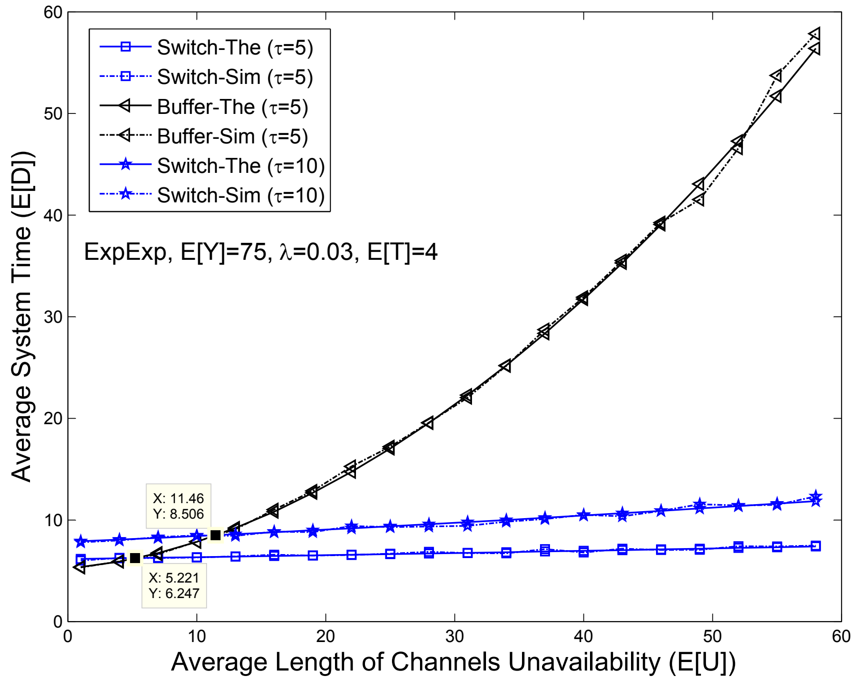

Figure 5.6 Decision on employing a buffering or switching policy fulfilled with analytical queueing results.

given by the value $E[R]$ where the average interruption time for both policies are equal, but is obtained by finding the value of $E[R]$ where (5.37) is the same for both OSA policies. Based on the knowledge of the CR users sensing and switching time, and the estimated values of the average channel availability and interruption period lengths Gabran *et al.* (2013), the OSA network can therefore use (5.37) to optimally decide between the switching and buffering policies to minimize the packets sojourn time. In the remainder of the paper, we will derive similar relationships that can be used to analyze and optimally control on OSA network with multiple classes of traffic with priority queueing differentiated services.

5.4 Priority Queueing

We can now tackle the analysis of the four priority queueing disciplines for the general queueing model with N classes of CR traffic. Let $\rho_b = \sum_{j=1}^{N} \rho_{b,j} = \sum_{j=1}^{N} \lambda_j E[X_{b,j}]$ and $\mathbb{A} = \min\{A_1, ..., A_N\} \rightarrow \mathbb{A} \sim EXP(\lambda)$. We also use the notation $P_{ae}(\lambda)$ and $R_r(\mathbb{A})$ to highlight that P_{ae} and R_r in (5.3) and (5.17), respectively, should be calculated with combined λ and $\mathbb{A}$. For the non-preemptive and preemptive priority queueing disciplines presented in Section 5.4.1 and 5.4.3, respectively, results for a general distribution for the operating periods are presented. For the exceptional non-preemptive and preemption in case of failure service disciplines, introduced in this paper and presented in Section 5.4.2 and 5.4.4 respectively, we only analyze the case of exponential operating periods due to the analytical complexity of those schemes without the assumption of memoryless operating periods. We also present in Section 5.5 an alternative approach to analyze the preemptive and non-preemptive disciplines for exponential operating periods.

5.4.1 Non-preemptive

Since the packet service can not be preempted in this scheme, the completion time of any packet for the three cases ('a','b' and 'u') will be the same as for the single traffic queue. The moments of the general completion time X_i for class of traffic i, $i = 1, \ldots, N$, can then be found by solving the system of N equations and N unknowns obtained from (5.5) for the N classes of traffic where ρ is replaced by $\sum_{j=1}^{N} \lambda_j E[X_j]$. Then, similar to an M/G/1 queue Bertsekas et Gallager (1992), we have :

$$E[W_i] = \frac{E[J]}{(1 - \sum_{j=1}^{i} \rho_{b,j})(1 - \sum_{j=1}^{i-1} \rho_{b,j})}. \tag{5.39}$$

where J is the remaining completion time of the packet in service and is given by :

$$E[J] = \sum_{j=1}^{N} \frac{\lambda_j}{2} E[X_j^2]. \tag{5.40}$$

Note that since no knowledge is available about the packet in service, the general completion time is used. However, the denominator represents the completion time of the queued packets which is $X_{b,j}$ for class j.

When Y is exponentially distributed, we have that the completion time for the three cases ('a','b' and 'u') has the same distribution in the alternative model presented in Section 5.3.4. Using the same approach as for (5.37), a closed-form relation can be obtained for the system time by using a queue model with an exceptional completion time X_e for the first packet which initiates a busy period Takagi (1991). X_e is given by :

$$X_e = \sum_{i=1}^{N} \frac{\lambda_i}{\lambda} X_{e,i} = \sum_{i=1}^{N} \frac{\lambda_i}{\lambda} \left[P_{ae}(\lambda) X_{b,i} + (1 - P_{ae}(\lambda))(X_{b,i} + R_r(A))\right]. \tag{5.41}$$

It is straightforward to find the first two moments of $X_{e,i}$ and X_e. We then obtain Takagi

(1991) :

$$E[D_i] = \frac{(1-\rho_b)(E[X_{e,i}] + E[X_{b,i}]) + \lambda E[X_{e,i}]E[X_{b,i}]}{1 + \lambda E[X_e] - \rho_b}$$
$$+ \frac{\lambda[(1-\rho_b)E[X_e^2] + E[X_e](\sum_{j=1}^{N} \lambda_j E[(X_{b,j})^2])]}{2(1 + \lambda E[X_e] - \rho_b)(1 - \sum_{j=1}^{i} \rho_{b,j})(1 - \sum_{j=1}^{i-1} \rho_{b,j})}. \tag{5.42}$$

The first term represents the average completion time and the second term, the average waiting time. Similar to (5.38), P_0, the steady-state probability of the system being empty, is given for this queue by :

$$P_0 = \frac{1-\rho_b}{1-\rho_b+\lambda E[X_e]}. \tag{5.43}$$

5.4.2 Exceptional non-preemptive

In this scheme, a packet which arrives in an empty unavailable system (case 'u') can be preempted at the end of the arrival recovery period by a higher priority packet which also arrives in the same recovery period. However, this is in fact a non-preemptive discipline for the alternative model presented in Section 5.3.4 since in this model a packet which arrives in an empty unavailable system does not start the service, but is queued waiting to obtain the server which will be given to the queued packet with the highest priority. We thus have a non-preemptive queue with initial setup time Takagi (1991). Using the same approach as in Section 5.3.5, we obtain that :

$$E[D_i] = E[X_{b,i}] + \frac{\sum_{j=1}^{N} \lambda_j E[(X_{b,j})^2]}{2(1 - \sum_{j=1}^{i} \rho_{b,j})(1 - \sum_{j=1}^{i-1} \rho_{b,j})}$$
$$+ \frac{(1-\rho_b)(\lambda E[S^2] + 2E[S])}{2(1 + \lambda E[S] - \rho)(1 - \sum_{j=1}^{i} \rho_{b,j})(1 - \sum_{j=1}^{i-1} \rho_{b,j})}. \tag{5.44}$$

The probability of the system being empty is equal to :

$$P_0 = \frac{1-\rho_b}{1+\lambda E[S]}. \tag{5.45}$$

5.4.3 Preemptive

In this scheme, the highest class is not affected by the other classes of traffic. Its completion time and system time can thus directly be found using the results presented in Section 5.3. Let us now analyze the performance of the low priority class for a two priority class system.

To solve this system, let us find the distribution of Y_2 and R_2, respectively the operating and interruption periods, from the perspective of the low priority (LP) packets . The system is unavailable for LP traffic both due to the activity of high priority (HP) users and due to channel interruptions. As illustrated in Fig. 5.7.a, Y_2 is the minimum between the time to the next interruption and the arrival of an HP packet : any one which arrives sooner initiates an interruption period for LP packets. We thus have :

$$Y_2 = \min(Y, A_1) \rightarrow 1 - F_{Y_2}(t) = (1 - F_Y(t))(1 - F_{A_1}(t)). \tag{5.46}$$

When Y is exponentially distributed with parameter α, the distribution of Y_2 can be given by :

$$F_{Y_2}(t) = 1 - e^{-(\lambda_1+\alpha)t}. \tag{5.47}$$

To calculate R_2, we have to distinguish between the events that caused the period of interruption. If, as illustrated in Fig. 5.7.a, a high priority (HP) packet arrived and preempted the low priority (LP) traffic, the length of R_2 is equal to one busy period of HP packets which is distributed according to $B_{b,1}$. On the other hand, if the channel interruption caused the unavailability, the two cases shown in Fig. 5.7.b and 5.7.c may happen. First, if no HP packet arrives during R, the length of R_2 is equal to $R|R < A_1$. If an HP packet arrives during R, R_2 will be $A_1|(R > A_1)$, the interruption period until the HP packet arrival, extended with an HP busy period B_{R_r}. B_{R_r} can be found from (5.27), replacing X_e by $X_{b,1} + R_r$ where $R_r = R - A_1|(R > A_1)$ is the remaining time of the server interruption after the HP packet arrival and X_b by $X_{b,1}$. We then have :

$$R_2 = \begin{cases} B_{b,1} & Pr(A_1 < Y), \\ R|(R \leq A_1) & Pr(Y \leq A_1 \& R \leq A_1), \\ A_1|(R > A_1) + B_{R_r} & Pr(Y \leq A_1 \& R > A_1). \end{cases} \tag{5.48}$$

Please be aware that in order to simplify the notation, from now on, we use the notation $V^{<Z}$ to denote $V|(V < Z)$ for any two random variables V and Z. The probability of an HP arrival during R can be given by :

$$P_{A_1<R} = \int_0^\infty (1 - e^{-\lambda_1 r})dF_R(r). \tag{5.49}$$

For exponential Y, we have :

$$Pr(Y \leq A_1) = \frac{\alpha}{\alpha + \lambda_1}, \tag{5.50}$$

They are here two new random variables. Not to be mistaken with identical instances of Y and R in previous sections.

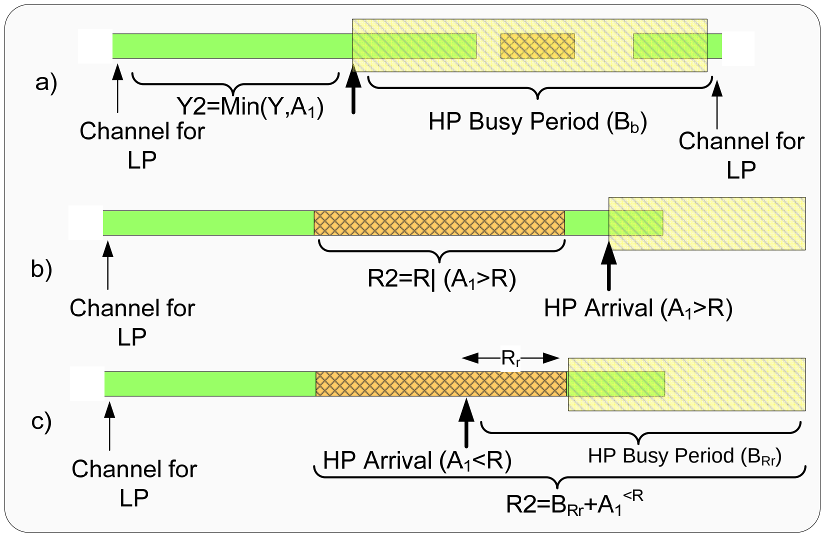

Figure 5.7 Operating and interruption periods (Y_2 and R_2) from the perspective of LP packets.

and then

$$E[R_2] = \frac{\lambda_1}{\alpha+\lambda_1}E[B_b] + \frac{\alpha}{\alpha+\lambda_1}\left[(1-P_{A_1<R})\frac{-d/d\lambda_1\widehat{R}(\lambda_1)}{\widehat{R}(\lambda_1)} + P_{A_1<R}(E[A_1^{<R}]+E[B_{R_r}])\right]. \tag{5.51}$$

Using (5.23) for $E[R_r]$ and (5.30) for busy periods, we have :

$$E[B_{R_r}] = \frac{E[X_{b,1}]+E[R_r]}{1-\lambda_1 E[X_{b,1}]} = \frac{E[X_{b,1}] + \frac{E[R]}{1-\widehat{R}(\lambda_1)} - \frac{1}{\lambda_1}}{1-\lambda_1 E[X_{b,1}]}. \tag{5.52}$$

The second moment of R_2 can be computed similarly, where the second moment of the busy periods can be found from (5.31) and the second moment of $A_1^{<R}$ and $R^{<A_1}$ can be derived from (5.22) in Section 5.3.1. It should be taken into account that B_{R_r} and $A_1^{<R}$ are correlated, so $E[B_{R_r}A_1^{<R}]$ should be calculated separately, using, for instance, the same approach as in (5.15).

From the equations above, one can find the moments of the operating and interruption periods from the perspective of LP packets (Y_2 and R_2). Then, we return to the original M/G/1 queue with interruptions and replace Y and R in (5.37) with Y_2 and R_2, respectively, to find the performance metrics of the LP packets.

Note that it is not easy to extend the proposed approach for more than two classes of CR traffic ; however, it can be used to find a bound for the performance of aggregated low priority traffic (combination of all low priority classes).

5.4.4 Preemption in case of failure

In this priority queueing model, high priority packets can only preempt the service from a low priority packet if an interruption occurs. When the service is resumed after an interruption, the priority is given first to high priority packets (HP). In other words, if an HP packet arrives while a low priority (LP) packet is in service, the HP service is started either after the end of the LP service or after an interruption, any one which occurs sooner. As expected, for any class of traffic the performance metrics for this scheme lie between the non-preemptive and preemptive schemes. The completion time of HP packets is not affected by this service discipline and can be found from the original single traffic queue. However, the waiting time of the HP packet is affected since it must wait until the end of the LP packet transmission or an interruption before starting its service. In the following, we first analyze the completing time of LP packets with this service discipline and then study the HP and LP waiting time. Finally, we discuss the special case where the service time of low priority packets is exponentially distributed.

Completion time of the LP packets

To find the completion time of the LP packets, we follow a similar approach as the one used for the preemptive scheme. Due to the memoryless operating periods distribution, from the LP packets perspective, the operating period distribution is not affected by the preemptive in case of failure service discipline and we have $Y_2 = Y$. On the other hand, R_2, the length of the interruption period from the LP users perspective, is a function of the remaining service time of the LP packet at the HP packet arrival time. That is, the longer the remaining service time until the next interruption, the more HP packets can arrive and thus their busy period will get longer. However, unless the service time is memoryless (this special case is discussed in Section 5.4.4), the remaining service time is not the same for each interruption R_2. Therefore, the completion time can not be modeled as a renewal process because the instances of R_2 are not identical. We will thus provide approximations for the moments of X_2 (or X_2^*) for two extreme cases : when the operating periods are much larger than the service time of type-2 (LP) packets ($Y >> T_2$) (*large scenarios*) and when it is smaller ($Y < T_2$) (*small scenarios*).

For $Y >> T_2$, it can be assumed that the service of an LP packet is finished in at most two type-2 operating periods. This assumption is a trade-off between accuracy and complexity and it is equivalent to assuming at most one instance of R_2 interruption during the completion time. The completion time will be found based on the alternate model (Section 5.3.4). The expectation of X_2^* is then given by :

$$E[X_2^*] \approx \begin{cases} E[T_2] & Y \geq T_2 \\ E[T_2] + E[R_2] & Y < T_2. \end{cases} \tag{5.53}$$

The length of R_2 depends on the arrival of an HP packet and its arrival time. R_2 can be given by :

$$R_2 = \begin{cases} R^{<A_1} & \text{No HP arrival,} \\ B_{C_r} - (Y^{<T_2} - A_1 | (A_1 < Y^{<T_2})) & \text{HP arr. in } Y^{<T_2}, \\ A_1^{<R} + B_{R_r} & \text{HP arr. in } R. \end{cases} \tag{5.54}$$

where C_r is the remaining time of the cycle (a cycle consists of an operating period Y followed by a recovery period R) after the arrival of an HP packet, and B_{C_r} represents the HP busy period which is initiated with $C_r + X_{b,1}$. However, $Y^{<T_2} - A_1$, the remaining time of the operating period $Y^{<T_2}$ after the HP packet arrival, should be excluded from R_2 since the HP packet does not immediately preempt the LP packet. In the third case, the HP arrival occurs in R. The busy period of HP packets thus starts with $R_r + X_{b,1}$ and the length of the total interruption is A_1 in addition to the HP busy period. $E[R_2]$ can thus be given by :

$$\begin{aligned} E[R_2] = {} & (1 - P_{A_1 < C^{<T_2}}) \frac{-d/d\lambda_1 \widehat{R}(\lambda_1)}{\widehat{R}(\lambda_1)} + P_{A_1 < C^{<T_2}} \\ & \Bigg[P_{ae}(\lambda_1) \Bigg(\frac{E[R] + E[Y^{<T_2} - A_1 | A_1 < Y^{<T_2}] + E[X_{b,1}]}{(1 - \lambda_1 E[X_{b,1}])} \\ & - E[Y^{<T_2} - A_1 | A_1 < Y^{<T_2}] \Bigg) + (1 - P_{ae}(\lambda_1)) \Bigg(E[A_1^{<R}] + \frac{E[X_{b,1}] + E[R_r]}{(1 - \lambda_1 E[X_{b,1}])} \Bigg) \Bigg], \end{aligned} \tag{5.55}$$

where $P_{A_1 < C^{<T_2}}$ is the probability of an arrival in $C^{<T_2} = Y^{<T_2} + R$, and P_{ae} is calculated for HP packets. The second moment of R_2 can be found similarly using the second moment of the busy periods and the relations provided in Section 5.3.1. However, the correlation of random variables B_{R_r} and $A_1^{<R}$ should be taken into account.

For the case where $Y < T_2$, we assume that the duration of HP busy periods is independent of the activity of LP packets. Therefore, the interruption periods from the perspective of LP packets have the same distribution and a renewal process can be considered to analyze the completion time. When a LP packet starts its service, as illustrated in Fig. 5.8, it holds the channel (available or unavailable) for a cycle and if there is an HP arrival during the cycle, it releases the channel to HP packets at the end of the cycle. So, the probability of releasing the channel to HP traffic at the end of the cycle is given by :

$$P_{A_1<C} = Pr(\text{HP arrival in } C) = \int_0^\infty (1 - e^{-\lambda_1 c}) f_C(c) dc. \tag{5.56}$$

For R_2, (5.54) is still valid except that we eliminate the condition that $Y < T2$.

We then obtain :

$$\begin{aligned} E[R_2] &= (1 - P_{A_1<C}) \frac{-d/d\lambda_1 \widehat{R}(\lambda_1)}{\widehat{R}(\lambda_1)} \\ &+ P_{A_1<C} \left[P_{ae}(\lambda_1) \left(\frac{E[R] + E[Y]}{(1 - \lambda_1 E[X_{b,1}])} - E[Y] \right) + (1 - P_{ae}(\lambda_1)) \left(E[A_1^{<R}] + \frac{E[X_{b,1}] + E[R_r]}{(1 - \lambda_1 E[X_{b,1}])} \right) \right]. \end{aligned} \tag{5.57}$$

The second moment can be found similarly. The moments of R_2 can then be substituted into the results for the single traffic queue to find the moments of the completion time of LP packets $E[X_2]$ and $E[X_2^2]$.

Waiting time

For the high priority (HP) traffic, the waiting time of HP arrivals during Y in a system empty of HP packets is affected by the presence of LP packets. For those packets, the new waiting time is zero if the LP queue is empty or, if the LP queue is not empty, the minimum of the remaining time of the arrival cycle and the remaining service time of the LP packet in service. The difficulty to compute the waiting time is thus the dependency of both types of traffic on each other. That is, while the waiting time of the HP packets is affected by the lower class, both the waiting time and the completion time of LP packets are affected by HP traffic. This obliges us to use approximations and bounds to find the waiting time of LP and HP packets.

The waiting time of HP packets is upper bounded by an M/G/1 queue with vacation in (5.37) if we neglect the unknown part of the service that the LP has received so far and assume that the remaining LP service time is still T_2. That is, the initial setup time S can be approximated as follows :

$$S = \begin{cases} R_r(\lambda_1) & 1 - P_{ae}(\lambda_1), \\ Y^{<T_2} + R & Pr(Y \leq T_2)(1 - P_0) P_{ae}(\lambda_1), \\ T_2^{<Y} & Pr(Y > T_2)(1 - P_0) P_{ae}(\lambda_1), \\ 0 & P_0 P_{ae}(\lambda_1), \end{cases} \tag{5.58}$$

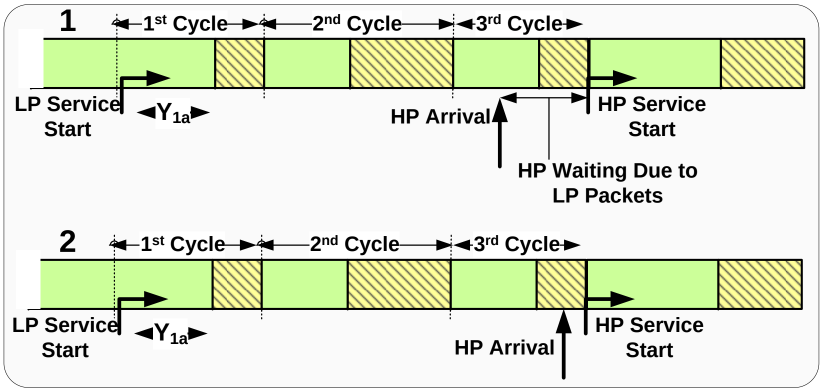

Figure 5.8 Cycles and holding periods for the LP packets in the discipline of preemptive in case of failure (FP). The LP packet can hold the channel for three cycles. However in a preemptive scheme (Pr) (not shown here), it can keep the channel for two complete cycles and releases the channel in the middle of the third cycle if there is an HP arrival (section 1).

where $P_{ae}(\lambda_1)$ and $R_r(\lambda_1)$ only take HP packets into account, and P_0 is the probability of system being empty of any type of packet. Note that the correct probability to be used here instead of $1 - P_0$ is $P_{L|NH}$, which is the probability that there are LP packets in the system given that it is empty of HP packets. However, this probability can not be found without any assumption on T_2's distribution ; therefore, we used P_0 as an approximation. This approximation results in a setup time, S, larger than its real value, which provides an upper bound in (5.59) that can be sometimes looser than expected. P_0 is the same for all priority queueing models and is given in (5.43). An obvious lower bound on the waiting time is given by assuming that HP packets always preempt LP packets (preemptive discipline). We can thus write :

$$\frac{\lambda E[X_{b,1}^2]}{2(1-\lambda E[X_{b,1}])} + \frac{E[R^2]}{2(E[Y]+E[R])} \leq E[W_1^*] < \frac{\lambda E[X_{b,1}^2]}{2(1-\lambda E[X_{b,1}])} + \frac{2E[S]+\lambda_1 E[S^2]}{2(1+\lambda E[S])}. \quad (5.59)$$

We can then use those bounds on the HP waiting time to find corresponding bounds on the LP waiting time through the Conservation Law (CL) in a queue with multiple classes of traffic Kleinrock (1975) which indicates that the quantities

$$\kappa = \lambda_1 E[T_1]E[W_1] + \lambda_2 E[T_2]E[W_2], \quad (5.60)$$

and

$$\kappa^* = \lambda_1 E[T_1]E[W_1^*] + \lambda_2 E[T_2]E[W_2^*], \quad (5.61)$$

for alternative model, are constant for all priority service disciplines. κ and κ^* can thus be computed with the waiting time of LP and HP packets found for one of the previous priority disciplines.

An alternative approach is to directly find the LP waiting time and then use the conservation law to obtain the HP waiting time. But, similarly to the HP waiting time, it is difficult to find an exact expression for the LP waiting time due to the strong interdependence between both types of traffic. We thus propose to compute bounds as follows. Using the approximations for the first two moments of X_2 or X_2^* found previously, the minimum waiting time of LP packets can be found using the Pollaczek—Khinchine (P-K) relation. An upper bound for the LP packets' waiting time is naturally given by the waiting time in the preemptive discipline model.

We thus have two upper and lower bounds for both the HP and LP waiting time. The tighter bounds can then be selected as the final lower and upper bound for both traffic categories.

Exponentially distributed LP service time

The queue model for the HP traffic is an M/G/1 queue with vacations. Since for a memoryless exponentially distributed service time, the remaining parts of the service are identically distributed, it is possible to exactly express S, the initial setup time for the HP traffic, as follows :

$$S = \begin{cases} R_r(\lambda_1) & 1 - P_{ae}(\lambda_1), \\ Y^{<T_2} + R & Pr(Y \leq T_2) P_{L|NH} P_{ae}(\lambda_1), \\ T_2^{<Y} & Pr(Y > T_2) P_{L|NH} P_{ae}(\lambda_1), \\ 0 & Otherwise, \end{cases} \tag{5.62}$$

where $P_{ae}(\lambda_1)$ and $R_r(\lambda_1)$ only take HP packets into account. $E[S]$ is then given by :

$$E[S] = (1 - P_{ae}(\lambda_1))E[R_r] + P_{L|NH} P_{ae}(\lambda_1) \left[\left(\frac{1}{\gamma_2 + \alpha} + \frac{\alpha}{\alpha + \gamma_2} E[R] \right) \right]. \tag{5.63}$$

α is the exponential parameter for Y and γ_2 is the exponential parameter for T_2. The unknown in the preceding equation is $P_{L|NH}$, which is the probability that there are LP packets in the system given that it is empty of HP packets, and can be expressed as :

$$P_{L|NH} = 1 - \frac{P_0}{P_{0,1}}, \tag{5.64}$$

where $P_{0,1}$, the probability that the system is empty of HP packets, can be found from the original queue by substituting $E[S]$ with the one that was calculated in (5.63), and, as indicated previously, P_0 is the probability that the system is empty of all packets and is given in (5.43). We thus have two equations ((5.63) and (5.64)) which can be used to find the two unknowns $E[S]$ and $P_{L|NH}$. The second moment of S can then be found and the HP waiting time is given by (5.37).

The LP packets waiting time can be found easily from the conservation law and the previous result for the HP packets waiting time. To find the system time, we will now find an exact expression for the first two moments of the completion time of LP packets. The interruptions, from the LP packets point of view, are identically distributed due to the exponential LP service time distribution. The same renewal process approach which was used as an approximation for $Y < T_2$ duration in (5.57) can thus be used exactly for this case. The only change is that Y should be replaced with $Y^{<T_2} = Y|(Y < T_2)$. We then have

$$\begin{aligned} E[R] = {} & (1 - P_{A_1<Y|Y<T_2} - P_{A_1<R})E[R^{<A_1}] \\ & + P_{A_1<Y|Y<T_2}(E[B_{C_r}] - E[Y_r^{<T}]) + P_{A_1<R}(E[B_{R_r}] + E[A_1^{<R}]), \end{aligned} \tag{5.65}$$

where

$$P_{A_1<Y|Y<T_2} = \frac{\lambda_1}{\lambda_1 + \alpha + \gamma}, \tag{5.66}$$

$$P_{A_1<R} = (1 - P_{A_1<Y|Y<T_2})Pr(R < A_1), \tag{5.67}$$

$$E[B_{C_r}] = \frac{E[R] + \frac{1}{\gamma_2+\alpha} + E[X_{b,1}]}{1 - \lambda_1 E[X_{b,1}]}, \tag{5.68}$$

$$E[B_{R_r}] = \frac{E[R_r] + E[X_{b,1}]}{1 - \lambda_1 E[X_{b,1}]}. \tag{5.69}$$

Similarly, $E[R_2^2]$ and consequently $E[X_2]$ and $E[X_2^2]$ can be found.

It is worth noting that only the assumption of exponential service time for LP packets is required to find the results above. The service time of HP packets can be general.

5.5 Alternative approach for priority queueing analysis

In this section, we propose an alternative approach to analyze the preemptive and non-preemptive priority queueing service disciplines in OSA networks. As interruptions have a preemptive behavior, we can model the interruptions as the highest priority type of traffic whose packet inter-arrival time is distributed with a random variable Y and whose *busy periods* are distributed with a random variable R. We can then have an estimate for the service time of these virtual highest priority packets whose busy period models the interruptions. Closed-form relations can not be derived in general since for any distribution of R and Y, a different formula for the busy periods and consequently for the service time exists. If Y is exponentially distributed with parameter α, we can assume an M/G/1 queue for the interruptions. Note that, as discussed in the introduction, this *alternative approach provides an approximation* because we know that the real distribution of the busy periods in an M/G/1 queue is a complicated function built on the Bessel function Takagi (1991), which can not exactly be matched to R. We discussed this alternative approach in this paper for completeness and as an extension of the analysis provided in Wang *et al.* (2011); Rashid *et al.* (2007); Laourine *et al.* (2010) for multiple classes of traffic with preemptive and non-preemptive service disciplines.

Using this alternative approach, it is possible to find approximate results for the preemptive and non-preemptive service disciplines as follows. We first find the first two moments of the service time of the virtual packets which form the interruptions from the distribution of the busy periods of a regular M/G/1 queue Takagi (1991) :

$$\widehat{R}(s) = \widehat{T}_v(s + \alpha - \alpha\widehat{R}(s)), \tag{5.70}$$

$$E[T_v] = \frac{E[R]}{1 + \alpha E[R]} \ , \ E[T_v^2] = E[R^2](1 - \alpha E[T_v])^3. \tag{5.71}$$

where T_v stands for the service time of the virtual packets which form the interruptions. As interruptions have a preemptive behavior, we can model the preemptive service discipline as a preemptive-resume queue with $N+1$ classes of traffic, the highest priority packets being the virtual packets. The following extensions of the P-K formula for preemptive-resume schemes Takagi (1991) can then be used to find the average waiting time of other classes of traffic :

$$E[W_i] = \frac{E[J_i]}{(1 - \alpha E[T_v] - ... - \lambda_{i-1}E[T_{i-1}])(1 - \alpha E[T_v] - ... - \lambda_i E[T_{i-1}])}, \tag{5.72}$$

where $E[J_i]$ can be given by :

$$E[J_i] = \frac{1}{2}\alpha E[T_v^2] + \sum_{j=1}^{i} \frac{1}{2}\lambda_j(E[T_j^2]), \tag{5.73}$$

and the moments of T_v can be found from (5.71). The system time is then given by :

$$E[D_i] = \frac{E[T_i]}{1 - \rho_{i-1} - \cdots - \rho_1 - \rho_v} + E[W_i] \tag{5.74}$$

For the non-preemptive service discipline, the queue can be modeled as a priority queue where the highest class of traffic (virtual) behaves preemptively, but other classes behave non-preemptively. The extensions of the P-K formula given in (5.72) can be used to find the average waiting time of other classes of traffic where $E[J_i] = E[J]$ is the same for all priority classes and is equal to :

$$E[J] = \frac{1}{2}\alpha E[T_v^2] + \sum_{j=1}^{N} \frac{1}{2}\lambda_j(E[T_j^2]). \tag{5.75}$$

The system time can be written as :

$$E[D_i] = \frac{E[T_i]}{1 - \alpha E[T_v]} + E[W_i]. \tag{5.76}$$

5.6 Simulation Results

In this section, we validate the analytical results of the OSA networks priority queueing disciplines presented in this paper by comparing with system accurate Monte-Carlo simulation results. The presented results also give several insights on the performance of OSA networks with mixed traffic. We consider in the numerical evaluation a system with two classes of traffic : high priority (HP) and low priority (LP), also denoted as type-1 and type-2 packets, respectively. We assumed exponentially distributed operating periods and considered the cases of exponentially and constantly distributed interruption periods. The service time (packet length) is also assumed to have either an exponential or a constant distribution.

Unless mentioned otherwise, the HP arrival rate is assumed equal to 0.03 and the average real service times are $E[T_1] = 3$ and $E[T_2] = 5$ (the unit of time is irrelevant). As an example, if the unit of time is *millisecond* (ms) and the channel data rate is 4 Mbps, those service times represent 1500 bytes and 2500 bytes packets, respectively.

The duration of operating and interruption periods are selected to model two different scenarios. The first scenario is for an almost quasi-static cognitive radio network where $E[Y] \gg E[T]$ and used $E[Y] = 75$ and $E[R] = 15$. The second scenario is for a highly dynamic cognitive radio network Khalife *et al.* (2009) where $E[Y] < E[T]$ and used $E[Y] = 1$ and $E[R] = 0.2$. Note that the average server availability is the same for both scenarios, only the dynamics are different.

In the figures, the different distribution cases are denoted by 'ExpExp', 'ExpDet', 'DetExp' and 'DetDet', respectively for the distributions of T and R, as summarized in Table 5.1. The four service disciplines are denoted in the figures as 'Non' (non-preemptive), 'ENo' (exceptional non-preemptive), 'Pr' (preemptive) and 'FP' (preemption in case of failure). The suffix 'Sim' indicates the simulation results, 'The' corresponds to the analytical evaluation of the theoretical results presented in Section 5.4, and 'Alt' indicates the analytical evaluation of the theoretical results for the alternative approach presented in Section 5.5.

Table 5.1 Service time and recovery time distribution cases.

Service Time (T)	Exponential	Exponential	Constant	Constant
Recovery Time (R)	Exponential	Constant	Exponential	Constant
Notation	ExpExp	ExpDet	DetExp	DetDet

5.6.1 Exponential recovery and service time

The recovery time for a random channel selection recovery algorithm (i.e., the user senses a list of channels one by one until finding an available channel) or a slotted-Aloha competition with other CR users can be accurately modeled with an exponential distribution. Fig. 5.9 shows the HP and LP average system time respectively for this case. As expected, and can also be observed for all the presented results, the HP system time increases from the preemptive, preemptive in case of failure, exceptional non-preemptive and non-preemptive service disciplines, and the LP system time increases in the inverse order of service disciplines. The results also show that the simulation and theoretical analysis results closely match, which validates the priority queueing analysis.

Results are also presented for static and dynamic operating period scenarios. We can observe that for the same server availability ratio $E[Y]/E[R]$, the system time is worst for both classes of traffic and all service disciplines for the static scenario where $E[R]$ and $E[Y]$ are much larger than the service time. This is due to the fact that the long recovery periods in the static scenario have a severe impact on the OSA queue performance metrics for all traffic classes. To further investigate this important finding, in Figure 5.10 we present the system time as a function of $E[T_1]/E[R]$ for a server availability ratio $E[Y]/E[R]$ fixed to five. This figure clearly shows that the OSA network system time performance deteriorates as the system dynamic decreases (i.e., when $E[T_1]/E[R]$ decreases) with an inflexion point when the service time is approximately equal to the average interruption length. Furthermore, both HP and LP packets are similarly affected. That is, queueing disciplines can not protect HP traffic against long interruptions. This is expected since interruptions indeed preempt the server. Note that for a traditional OSA throughput analysis based on a saturated-traffic model, no major performance changes will be observed as a function of the OSA network dynamic since the main factor is the server availability ratio $E[Y]/E[R]$. Only the complete queueing analysis presented in this paper can give an insight on the important impact of system dynamics on the OSA performance.

The other interesting point to observe is that for dynamic scenarios, due to frequent short interruptions, the preemption in case of failure service discipline enables the quick

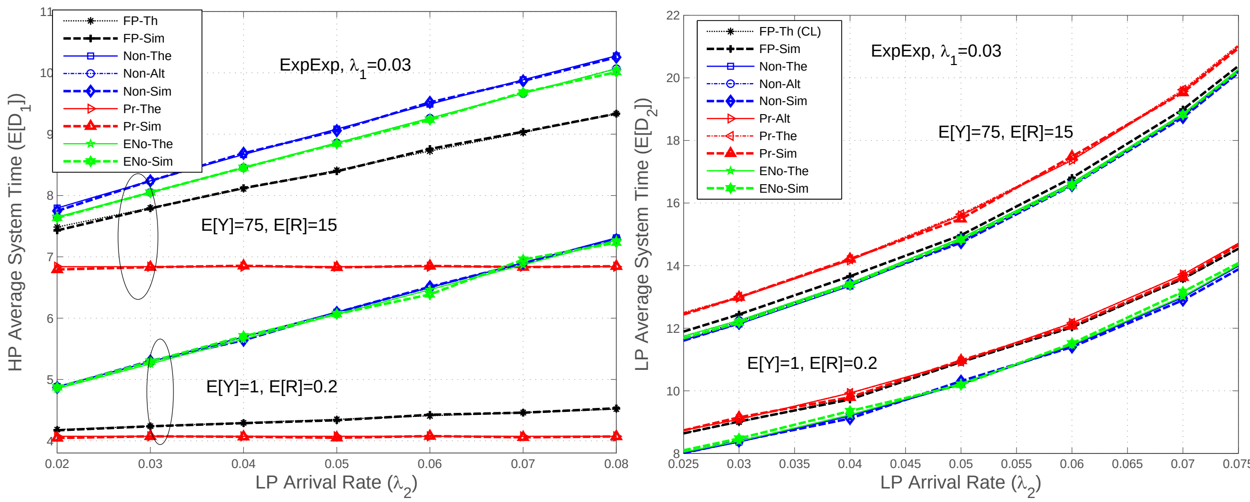

Figure 5.9 System time of high priority (HP) and low priority (LP) packets vs. LP arrival rate.

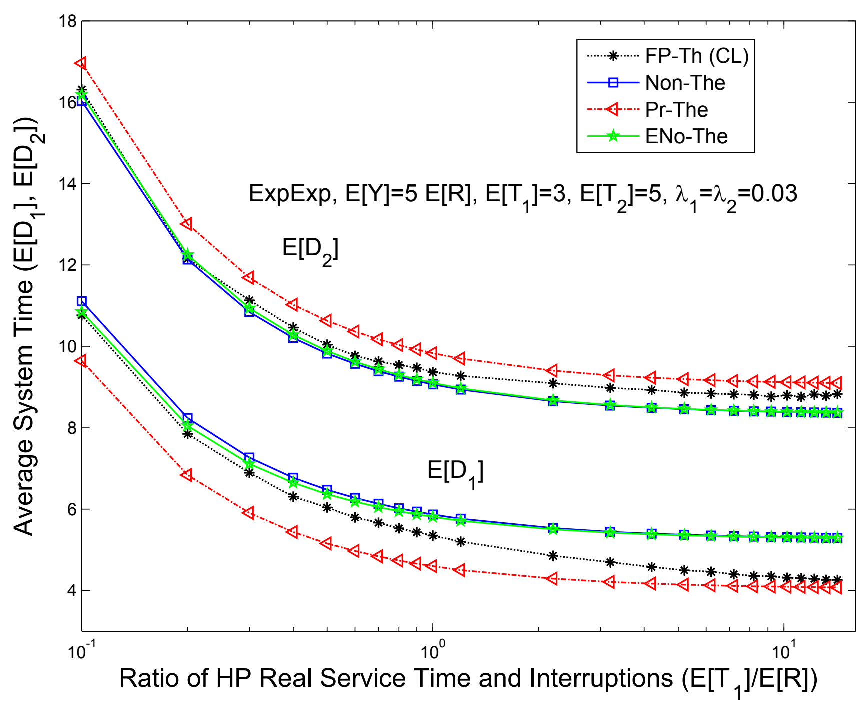

Figure 5.10 System time of LP and HP packets vs. the variations of $E[R]$ and $E[Y]$ when their ratio is fixed.

preemption of LP packets by HP packets. This service discipline performance is thus close to the preemptive scheme for dynamic scenarios. On the other hand, for static scenarios the preemption in case of failure service discipline performance gets closer to the non-preemptive discipline due to the lack of opportunities for HP packets to preempt LP packet service. Meanwhile, the system time for the exceptional non-preemptive discipline is very close to the non-preemptive scheme performance in dynamic scenarios because the probability of an HP arrival in an empty system in the same recovery period as an LP arrival is very low. In large scenarios, their performances start to differ. However, the performance gain remains small. Those results show that the novel priority discipline of preemption in case of failure for OSA networks can significantly improve the system time of HP packets in several deployment scenarios while having a lower implementation complexity than a full preemptive service discipline.

In Fig. 5.11, we study the CR traffic system time as a function of the operating and interruption period length. Those results illustrate the validity of the queue analysis for a wide range of operating and interruption periods. Note that since $E[R]$ or $E[Y]$ is fixed, the server availability increases as a function of $E[Y]$ in the former case and decreases as a function of $E[R]$ in the later case. We can observe that as the server availabitily decreases, either due to shorter availability periods or longer interruption periods, the system time significantly increases, with LP traffic being more affected than HP traffic due to the priority service disciplines. It is interesting to again note that the system time increases much faster when the interruption period increases than when the availability period decreases. For example, starting from the point where $E[Y] = 75$ and $E[R] = 15$ to the point where $E[Y] = E[R]$, the HP system time approximately increases by a factor of three when the operating period length decreases and by a factor of eight when the interruption period length increases. Those results underline the critical importance of minimizing the interruption period length in OSA networks.

The results presented in Fig. 5.12 further motivate the importance of the theoretical analysis provided in this paper to correctly analyze the performance of OSA networks. A straightforward tempting simplification that could be used to analyze the CR queue system time is to use the standard M/G/1 formulas without interruption and increase the packets real service time T_i by the ratio $\frac{E[Y]+E[R]}{E[Y]}$ to compensate for the average throughput loss due to interruptions. It can be shown for both traffic classes that the queues will saturate at the same traffic load for both the simplified analysis and the correct analysis. However, as can be seen in Fig. 5.12, the M/G/1 simplification (referred as 'NoInt') significantly underestimates by almost an order of magnitude the real performance of the queue for both preemptive and non-preemptive priority disciplines (this simplification does not allow the analysis of the two

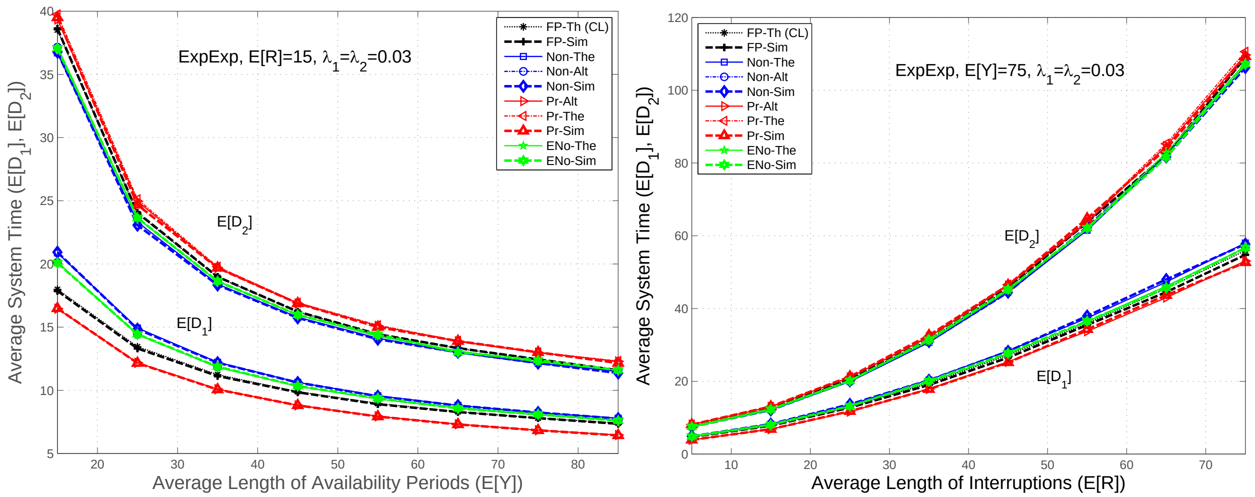

Figure 5.11 System time of HP and LP packets vs. operating and interruption period duration.

other OSA service disciplines due to the absence of interruptions). This error is due to the fact that the simplified modeling is in fact equivalent to assuming that the interruption periods approach a length of zero. But, as we have discussed previously, interruption periods have a major impact on the OSA queue performance. The accurate modeling of the interruption periods, as we provided in this paper, is thus critical to obtain a valid OSA queue analysis.

Fig. 5.12 also presents the system time when both traffic classes are mixed without a priority service discipline (i.e., both packet types are queued together and are served in a first-in-first-serve scheme). It is interesting to observe that for a standard system without interruption, the system time increase for LP packets when a priority service discipline is used is almost the same as the system time decrease for HP packets (i.e., the mixed traffic service time is almost exactly in the middle between the LP and HP service times with both priority service discipline). However, this is not the case for the OSA network where the HP packet system time decreases significantly more than the system time increase for LP packets. This is due to the fact that the interruption periods preempt both classes of traffic. Those results indicate that differentiated service with priority queueing has a bigger impact in OSA networks than in conventional networks and should thus be actively used when they carry multiple classes of traffic with different QoS requirements.

5.6.2 Exponential recovery time and constant service time

In Fig. 5.13, the HP and LP packet lengths are both constant ('DetExp' scenario). Thus, the lower bound for the waiting time of LP packets for the preemptive in case of failure service discipline is presented. The results are generated for two different values of HP arrival rate.

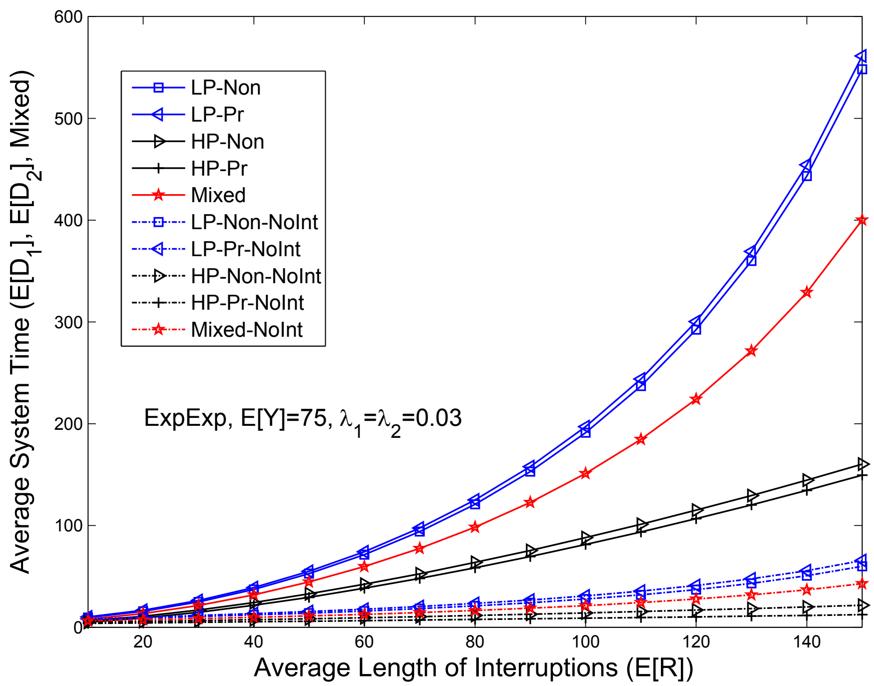

Figure 5.12 Performance comparison of accurate queueing models with interruption with a modified M/G/1 with longer packets but no interruption.

To better observe the accuracy of the completion time approximations for the preemptive in case of failure service discipline, simulation results and approximations for the moments of the LP packets completion time, X_2 and X_2^* , are compared in the upper part of Table 5.2. Note that the completion time of LP packets is independent of their arrival rate. The results show the accuracy of the analysis for non-exponential service time and the slight deterioration due to the approximate analysis for the preemptive in case of failure service discipline. On the other hand, we can observe the accuracy of the theoretical analysis with deterministic packet service time for the three other service disciplines.

5.6.3 Constant recovery time

A constant recovery time occurs, for instance, in scenarios where the information concerning the channels' occupancy is provided in advance ; therefore, no random sensing is required and the recovery time only represents a constant time for negotiation and radio alignment. In order to compare the results with the previous scenario, we assume the same average values.

Fig. 5.14 illustrates the system time of HP and LP packets versus their arrival rate for the cases of exponential and constant service times. It should be noted that for the selected values in the simulation, the upper and lower bounds proposed in Section 5.4.4 (e.g., Eq. (5.59))

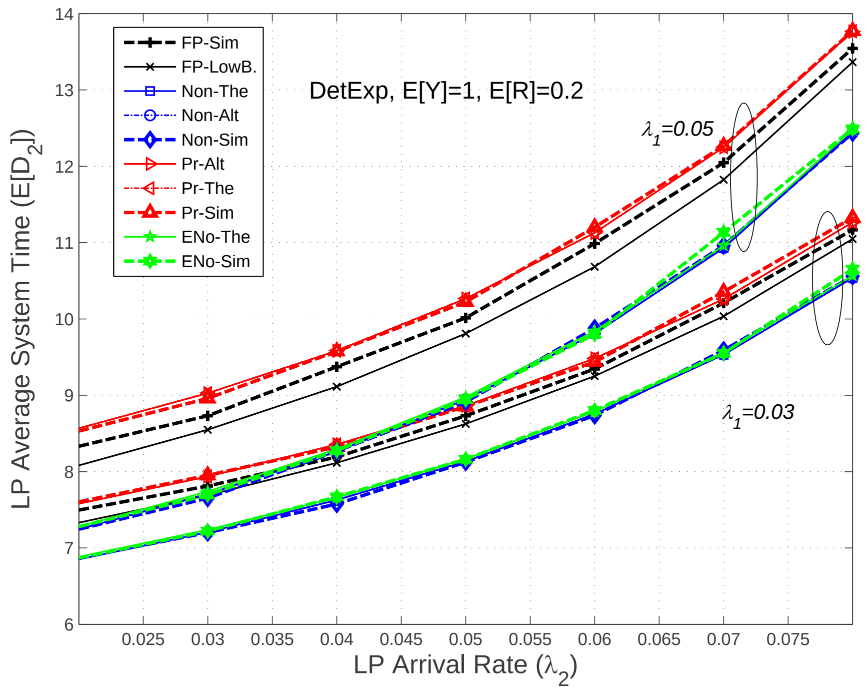

Figure 5.13 System time of LP packets vs. LP arrival rate for two values of HP arrival rate (small scenario).

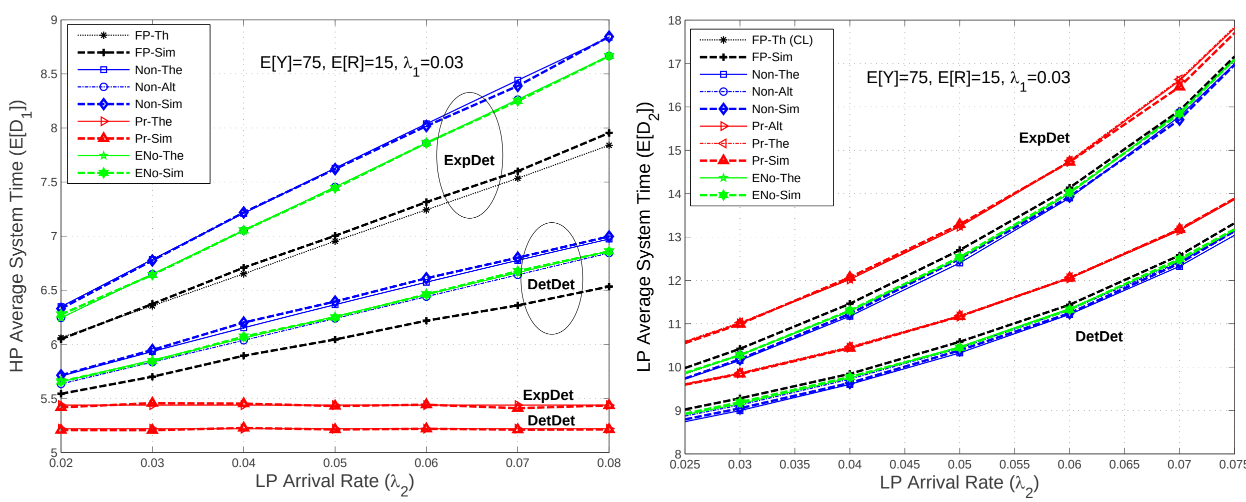

Figure 5.14 System time of high (HP) and low priority (LP) packets vs. LP arrival rate (DetDet and ExpDet, large scenario).

are loose compared to the natural bounds of preemptive and non-preemptive disciplines, so they are not illustrated in the figures to enhance their clarity. As expected, it can be seen in both figures that the performance, when the distribution of real service time (packet length) is exponential, is worse compared to the case where the packet length is constant (with the same average). The results also validate the theoretical analysis presented in Section 5.4. However, it can be observed that the system time for the non-preemptive service discipline obtained with the alternative approach presented in Section 5.5 is not accurate. This shows, as discussed in the introduction, the limitations of alternative approaches which were previously proposed in the literature when the recovery period is not exponentially distributed. Furthermore, this alternative approach can not be used to analyze more sophisticated service disciplines such as the exceptional non-preemptive and the preemptive in case of failure disciplines which provide interesting performance gains for OSA networks. Simulation and analytical results for the completion time of LP packets are compared in the lower part of Table 5.2 for two different values of HP arrival rate.

5.7 Conclusion and Future Work

Priority queueing is a classical approach to implement traffic differentiation in communication links. To analyze priority queueing schemes for opportunistic spectrum access networks, we derived in this paper a general queueing model with interruptions for the preemptive and non-preemptive classical priority disciplines. Two new cognitive radio disciplines were also introduced in this paper : exceptional non-preemptive and preemptive in case of failure. The theoretical analysis was validated with simulation results and we investigated the behavior of those disciplines for different sets of parameters and distributions for the packet service time and interruption periods. We also showed how the analysis can be used by an OSA controller to make critical decisions such as selecting the channel switching policy or the

Table 5.2 Moments of the LP completion time with preemption in case of failure (FP) service discipline for DetExp and DetDet scenarios (S : Simulation, A : Approximation).

Scen.	λ_1	$E[X_2]$-**S**	$E[X_2]$-**A**	$E[X_2^2]$-**S**	$E[X_2^2]$-**A**
Small,DE	0.03	6.62	6.72	47.01	49.51
Small,DE	0.05	7.10	7.32	57.05	62.33
Large,DE	0.03	6.13	6.12	77.53	74.96
Large,DE	0.05	6.26	6.25	88.37	85.62
Small,DD	0.03	6.61	6.72	46.76	49.23
Small,DD	0.05	7.10	7.31	56.69	61.98
Large,DD	0.03	6.14	6.12	58.40	56.45
Large,DD	0.05	6.26	6.25	65.00	63.64

priority queueing discipline based on the estimated channel parameters. It was also observed that even though the ratio of operating and interruption periods plays an important role, a significant performance decrease is observed in a semi-static network with long operating and interruption periods compared to a fast-varying network with short periods and the same ratio. A simplified M/G/1 model with no interruption and compensated increased packet length can not thus capture the queue metrics performance of OSA networks. We also presented results demonstrating the importance of priority queueing to provide differentiated service in the presence of frequent interruptions. As discussed in the introduction, an important area of future work is to use the results presented in this paper to further study and optimize MAC protocols and channel assignment policies in OSA networks based on not only a saturated mode throughput analysis but also on queue metrics. Another interesting area of research is to extend this work to the cases of queueing with service repeat after an interruption and for non-homogeneous channels with variable service rate for different operating periods.

5.8 Further discussions (Not a part of the paper)

All the parameters and notations used throughout the chapter are summarized in Table 5.3.

It is also worth noting that the unit of time is irrelevant and can be any appropriate time unit, e.g., millisecond (ms) or microsecond (us). In the simulation parts and unless otherwise mentioned, the time unit is millisecond (ms).

In Table 5.2, the time unit is millisecond and the objective is to show that the approximations are very close to simulation (accurate) results.

Table 5.3 Notations

Notation	**Description**
Y	Length of operating periods (RV)
R	Length of recovery (interruption) periods (RV)
C	C=Y+R
λ	Packet arrival rate
A	packet inter-arrival time (RV)
X	Completion time (RV)
X^*	Completion time in alternative model (RV) (Section 5.3.4)
Index $_b$	For packets entered a busy system
Index $_e$	For packets entered an empty system
Index $_a$	For packets entered an empty-available system
Index $_u$	For packets whose service started at the beginning of a Y
B	HP busy periods (RV)
B_b	Busy period started with X_b
B_Z	Busy period started with $Z + X_b$
T	Real service time (RV)
J	Remaining completion time of the packet in service
$\hat{Z}(s)$	LST of a continuous random variable Z
$f_Z(t)$	P.D.F of a random variable Z
$F_Z(t)$	C.D.F of a random variable Z
$m(t)$	Average number of renewals until time t
$m_{a\|b\|u}(t)$	m(t) for packets of type a, b or u
$m^2(t)$	Second moment of the number of renewals until time t
ρ	$\lambda E[X]$
ρ_b	$\lambda E[X_b]$
P_{ae}	Prob. of system being available when empty (arrival point)
W	Waiting time in the queue
W^*	Waiting time for the alternative model
D	Total time spent in the system
$P_{aiC\|Y\|R}$	Probability of HP arrival in a cycle (C), Y or R
K	Local parameter to count the number of an event
$i = 1, 2$	Subindex represents the traffic class
α	Exp. dist. parameter when $F_Y = 1 - e^{-\alpha t}$
β	Exp. dist. parameter when $F_R = 1 - e^{-\beta t}$
$\tilde{F}_Z()$.	Min. of random variable Z and an Exp. dist. (Eq. 5.4)
γ	Exp. Dist. parameter when $F_T = 1 - e^{-\gamma t}$
N_p	Number of priority classes
T_v	Service time of virtual packets which form the interruptions
$P_{A<C}$	Probability of arrival in C=$Pr(A < C)$
R_r	Remaining of R after an arrival in R
C_r	Remaining of C after an arrival in C
$Z_1^{<Z_2}$	$Z_1\|(Z_1 < Z_2)$ (for two RVs)
S	Initial setup time (RV) (Section 5.3.5)
P_0	Probability of system being empty
$P_{L\|NH}$	Probability of system being empty of HP but not LP

CHAPTER 6

ARTICLE 4 : QUEUEING MODEL FOR HETEROGENEOUS OPPORTUNISTIC SPECTRUM ACCESS

Arash Azarfar, Jean-François Frigon and Brunilde Sansò
Submitted to IET Communications, March 2014
Department of Electrical Engineering
École Polytechnique de Montréal
C.P. 6079, succ. centre-ville, Montréal, QC, Canada, H3C 3A7
Email : {arash.azarfar, j-f.frigon, brunilde.sanso}@polymtl.ca

In this paper, we propose a queueing model to analyze the performance of an opportunistic spectrum access (OSA) system with service interruptions operating over heterogeneous channels in which the service transmission rate and the service interruption rate after the transmission is resumed are generally different than their value prior to the interruption. We first propose Markov chain models to analyze this system under memoryless service time and availability periods. Based on simplification assumptions, we also provide an analytical z-Transform analysis of the Markov models. The Markov model and approximations are validated with accurate system simulations. We also provide numerical results illustrating the non-convex relations between the traffic metrics and system parameters and that the proposed models are essential for optimal OSA network planning and operation. We further analyze and discuss the OSA queuing model for general distribution of service time and availability periods. The analytical and simulation results indicate that for usual system parameters, the queue average occupancy is similar for different distributions of service time and availability periods and that the memoryless Markov models can be used to accurately predict the heterogeneous OSA system traffic performance.

6.1 Introduction

Opportunistic spectrum access (OSA) is expected to be widely deployed in next generation wireless networks to address the fast traffic growth in wireless networks Jondral (2007); Mitola (2009); Xiao *et al.* (2013); Cisco (2013). In OSA networks, the packet transmissions are frequently interrupted because the cognitive radio (CR) users must stop using their

AZARFAR, A., FRIGON, J.-F. and SANSO, B. (2014) Queueing model for heterogeneous opportunistic spectrum access. Submitted to IET Communications.

operating channel when the channel's primary users appear or if the channel quality drops below a minimum threshold. During the interruption, depending on the OSA medium access control (MAC) algorithm, the CR users might stay on the channel until it becomes available again or switch to a new channel Park *et al.* (2011). In both cases, the operating channel after the interruption might have different parameters (channel bandwidth, probability of primary user arrivals, propagation conditions, level of interference, etc.) than the operating channel prior to the interruption.

Our objective in this work is to obtain a queueing model, which is is an important tool to analyze CR traffic performance Wang *et al.* (2011), in an OSA network with heterogeneous operating channels. In the queueing model, the operating channel is the server of the queue. We therefore have a system with a server with a time-variant service rate subject to frequent service interruptions occurring at a time-variant rate.

6.1.1 Related Work

Queueing models for servers with interruptions have been previously studied Federgruen et Green (1986); Avi-Itzhak (1963); Takagi (1991); D. P. Gaver (1962); Fiems *et al.* (2008). However, all those works have considered a server with a time-invariant service rate and server interruption rate. Two-class preemptive queuing models have been proposed for OSA networks where the CR server interruptions are modeled as the busy periods of the preemptive primary traffic Wang *et al.* (2011); Rashid *et al.* (2007); Laourine *et al.* (2010). With this approach, the server is considered time-invariant with a constant service rate and interruption rate. Furthermore, it is not straightforward to use those models to study OSA networks with generally distributed operating and interruption periods.

In Azarfar *et al.* (2012a) we proposed a new queueing model for OSA networks for a single class of CR traffic for general operating and interruption period lengths. However, a time-invariant service rate and server interruption rate is also assumed in this model. In Su et Zhang (2008) an OSA queuing model with variable service rate is studied. But the work did not address the recovery periods and variable interruption rate. The queueing model is also specific to the MAC protocol studied in the paper. In Rashid *et al.* (2009), the authors discuss a queueing model for a multi-user cognitive radio network with a variable service rate, but the notion of recovery periods is not considered.

6.1.2 Contributions

Our major contribution in this paper is, to the best of our knowledge, the first accurate Markov chain (MC) queuing model for an OSA system using heterogeneous channels with different transmission and interruption rates in the presence of interruptions with random periods. We also introduce another MC model for generally distributed interruptions where the arrival process during interruptions is approximated as a bunch arrival. Those two models can be numerically solved to find the distribution of the number of packets and therefore analyze the OSA network traffic metrics for different network parameters' values.

In some cases, such as for large number of channels and queue lengths, or for real-time control algorithm implementations such as the MAC protocol, channel sensing order selection, and user admission, the MC models can be computationally too intensive to solve. Another important contribution of this paper is therefore a novel analytical z-transform analysis to approximate the traffic metrics for the heterogeneous OSA network queue model. The z-transform analysis is also extended to general distributions of service time and availability periods.

The queueing models developed in this paper can be used to analyze the traffic metrics of different OSA networks using a set of heterogeneous channels. Although the performance study of an OSA network for specific MAC protocols is outside the scope of this paper, we present numerical results in this paper illustrating the complex non-convex relationships between the traffic metrics and demonstrating that tools, such as the ones we introduce in this paper, are essential for OSA network planning.

The remainder of this paper is organized as follows. Section 6.2 presents the OSA system and queueing models considered in this paper. In Section 6.3, we present the accurate two-dimensional MC model and the approximate z-transform analysis for exponentially distributed recovery periods. The case of general interruption periods is studied in Section 6.4. The accuracy of those different models and approximations is verified with Monte-Carlo simulations of the system in Section 6.5. We further analyze and discuss in Section 6.6 the OSA queuing model for general distributions of service time and availability periods. Finally, Section 6.7 concludes the paper.

6.2 System Model

As illustrated in Figure 6.1, the CR node alternates between operating and recovery periods. The operating periods begin from the time that the CR node switches to a new channel and last until a recovery or spectrum search has to be performed. Without loss of generality, we will use the term recovery to designate the period during which the CR node must stop its transmission and search over channels according to an arbitrary channel search algorithm depending on the protocol used by the CR network and, possibly, compete with other users to reserve a channel according to a medium access control (MAC) protocol. At

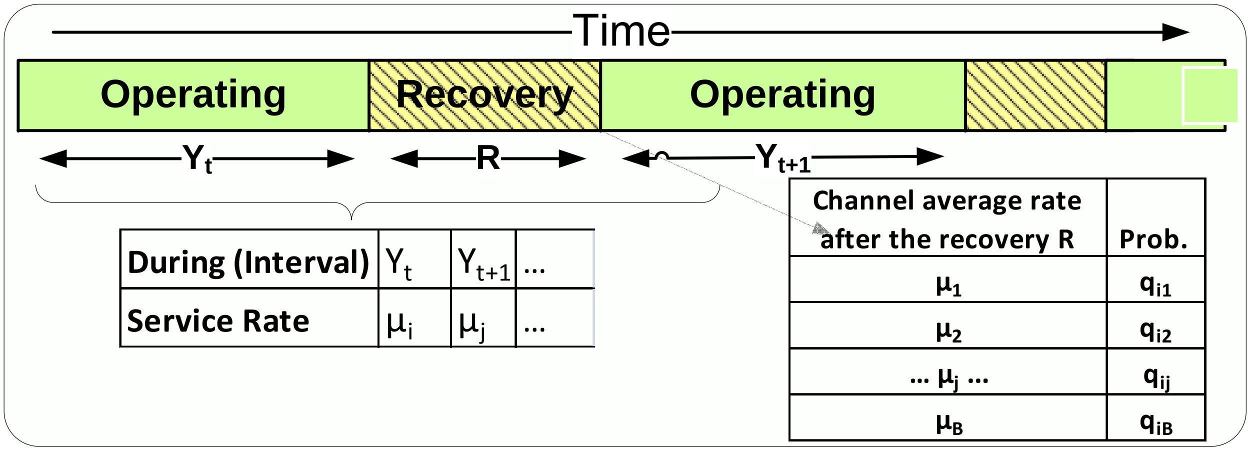

Figure 6.1 Cognitive radio node operation in opportunistic spectrum access.

the end of the recovery period, the CR node will select a channel of type i, $i \in \{1, \ldots, B\}$, and resume its packet transmission. We assume in this paper that the B types of channels are heterogeneous with different parameters such as channel bandwidth, link signal to noise plus interference ratio (SINR) and primary user occupancy distribution. Therefore, the system parameters are different between the operating periods before and after a recovery period. We also use, without loss of generality, the term failure event to designate the event triggering the start of a recovery period or equivalently the end of an operating period. A failure event can be due to several factors such as the appearance of primary users, a bad link quality in the operating channel or a periodic trigger.

The operating period has a random length Y with a general distribution. We assume that the distribution type is the same for the B types of channel (for example, they are all exponentially distributed) but the parameter (for example, the average) is different for each channel. The general distribution case is studied in Section 6.6.1 but, as discussed there, it is very difficult to obtain analytical queue performance results for non-memoryless operating period distributions. For this reason, the exponential distribution has widely been used in the literature to model the operating periods' length Yin *et al.* (2012); Su et Zhang (2008); Geirhofer *et al.* (2008); Huang *et al.* (2008). Results presented in this paper also show that the analytical results with the exponential distribution are good approximations for other distributions. For the memoryless case analyzed in Sections 6.3 and 6.4, the operating period over channel type i, $i \in \{1, \ldots, B\}$, is modeled with an exponential distribution with parameter α_i. Thus, when the CR node is operating on channel i, failure events occur at a rate α_i.

The service time depends on both the packet length and the channel transmission rate.

In this model, we consider a time-invariant Poisson packet arrival process with rate λ and time-invariant packet length distribution and parameters. During an operating period, it is also assumed that the channel transmission rate is constant. However, the transmission rate is channel dependent (for example, due to the different bandwidth or channel link quality) and changes for each operating period. We therefore have a system where the service time probability distribution is the same for all operating periods, but its parameters change for each operating period (for example, the service time is exponentially distributed with a different mean for each operating period). The general service time distribution case is studied in Section 6.6.2 and the memoryless packet length (and thus memoryless service time) is analyzed in Sections 6.3 and 6.4. For the later case, the service rate for channel type i, $i \in \{1, \ldots, B\}$, is modeled with an exponential distribution with parameter μ_i.

The recovery period had a random length R. Since the recovery period length depends on the system parameters, we model it with a time-invariant distribution which does not depend on the channel used during the previous operating period. In Section 6.3, we study the case of a recovery period with an exponential distribution with the parameter β (i.e., $E[R] = 1/\beta$) and in Section 6.4 we consider a generally distributed recovery period. The probability that channel j will be selected at the end of the recovery period given that channel i was used before the recovery period is assumed to be known and it is given by q_{ij}.

Several factors affect the distribution of the recovery period R and the channel selection probabilities q_{ij}. In multiuser scenarios, the MAC and opportunistic scheduling of users influence the distribution of R and the probabilities q_{ij}. For instance, in probabilistic MAC models, based on Aloha or CSMA, the recovery time to find a new channel includes not only the search time, but also the competition time (including backoff periods) with other nodes Park *et al.* (2011); Azarfar *et al.* (2014a). Even if the user is blocked due to other users transmitting on all channels, the total time of blocking until a successful channel reservation is included in the recovery time, which can be found for instance using a renewal model (i.e., if the user is blocked, the competition process to reserve a channel is renewed). In a network where a spectrum server assigns the channels, the recovery time is a deterministic short period of negotiation with the spectrum server until a new channel is assigned. Other factors, such as the channel search algorithm, channel sensing strategy, and number of channels should also be taken into account to determine the distribution of R. Note that the exponential distribution is a good approximation for the commonly used random channel search where channels are sensed one by one until the first available channel is found Luo et Roy (2007) and for multi-user Aloha competition for channel reservation Azarfar *et al.* (2014a). A methodology to find the recovery period distribution for two baseline multichannel opportunistic spectrum access MAC protocols is also provided in Azarfar *et al.* (2014a).

q_{ij}, the probability of finding a channel with a specific service rate after the period R, also depends on several factors such as the channel availability, channel models (fading and interference), multi-user channel assignment, and, in some cases, on the service rate of the channel used before the recovery. For example, we are using the common assumption in the literature Su et Zhang (2008); Park *et al.* (2011) that each channel is used by one user. However, given that multiple users may share the channels, we can assume that when a channel is multiplexed between multiple users, its service rate is also divided equally between the users. So, in the proposed queueing model, we can have a row for each possible service rate which would depend on the number of users assigned to the channel. The q_{ij} then depends on the number of users assigned to a channel and the channel scheduling strategy. Naturally, finding the q_{ij} values with channel sharing can be more involved Ma et Tsang (2008).

A combination of prior statistical knowledge about the channels (availability probability, channel fading model, etc.) in addition to system parameters (MAC protocol, channel search and sensing algorithms, channel assignment algorithm, number of users, number of channels, etc.) is therefore required to determine the distribution of R and the q_{ij}. The approach to find those is thus case-dependent and is out of the scope of this paper. However, the objective of this paper is to propose and analyze a general queueing model which can be used to find the packet level performance once those distributions are found. As discussed, the proposed model in this section is general and may be used for per-node performance analysis in different heterogeneous multi-channel opportunistic spectrum access network scenarios with multiple homogenous users.

6.3 Queue Model for Exponentially Distributed Recovery Periods

In this section, we consider exponential distributions for the operating periods, service time, and recovery periods, which enable the development of an accurate queue model. In Section 6.3.1, we present a Markov chain model for this CR queue and in Section 6.3.2, we derive approximate analytical queue performance results.

6.3.1 Markov Chain Model

Figure 6.2 shows the exact bi-dimensional Markov chain model for the case of exponentially distributed recovery periods. State (k, i), $k = 0, 1, \ldots,$ and $i = 1, \ldots, B$, indicates that there are k packets in the system operating over a channel of type i. State $(k, 0i)$, $k = 0, 1, \ldots,$ and $i = 1, \ldots, B$, indicates that there are k packets in the system which is in a recovery period, and the last operating channel before the start of the current recovery period has been channel i. When a failure occurs, the state makes a transition from (k, i) to state $(k, 0i)$,

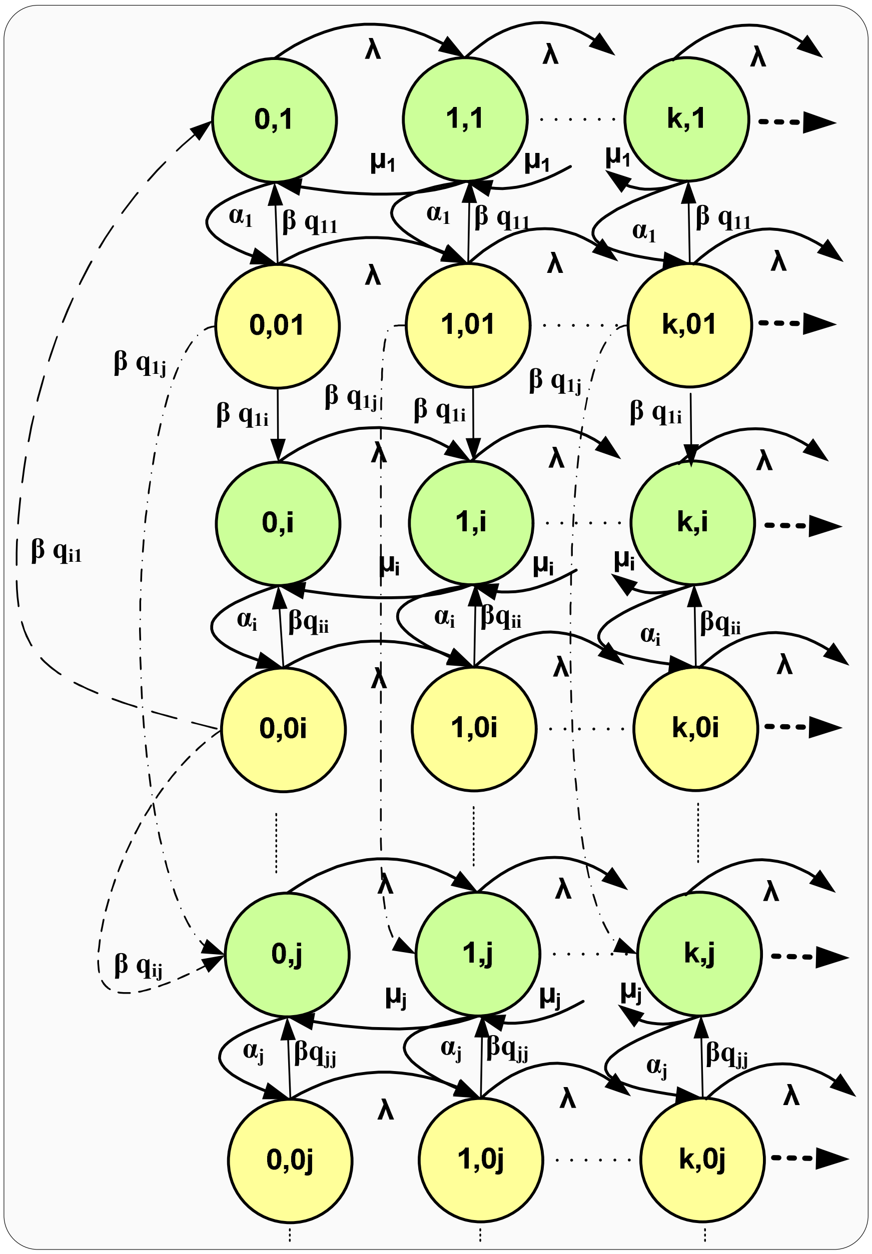

Figure 6.2 Markov chain for the queue with variable service rate and exponential recovery periods (M1).

while at the end of a recovery period, the state makes a transition from $(k', 0i)$ to state (k', j) with a probability q_{ij}. The states $(k, 0i)$ enable us to accurately model the packet arrivals during the recovery period, but the recovery periods must be exponentially distributed to be able to introduce them into the Markov chain. Note that B separate lines are needed for the recovery periods because of the memory between consecutive operating channels.

For convenience, let us define $\mu_{0i} = 0, \forall i > 0$. The transition rates are then as follows :

$$p_{(k,i)(l,j)} = \begin{cases} \lambda & i = j \text{ and } l = k+1 \\ \mu_i & i = j, i \in \{1, \ldots, B\} \ , \ l = k-1 \text{ and } k > 0 \\ \alpha_i & i \in \{1, \ldots, B\} \ , \ j = 0i \text{ and } l = k \\ \beta q_{mj} & i = 0m \ , \ m \in \{1, \ldots, B\} \ , \ j \in \{1, \ldots, B\} \text{ and } l = k \\ 0 & \text{Otherwise.} \end{cases} \tag{6.1}$$

The balance equations are then given by

$$\pi(0, i)(\alpha_i + \lambda) = \pi(1, i)\mu_i + \beta \sum_{j=1}^{B} \pi(0, 0j)q_{ji} \qquad i = 1, \ldots, B \tag{6.2}$$

$$\begin{aligned} \pi(k, i)(\alpha_i + \lambda + \mu_i) = \pi(k+1, i)\mu_i + \pi(k-1, i)\lambda& \\ + \beta \sum_{j=1}^{B} \pi(k, 0j)q_{ji} \qquad k > 0,\ i = 1, \ldots, B& \end{aligned} \tag{6.3}$$

and

$$\pi(k, 0i)(\beta + \lambda) = \pi(k-1, 0i)(\lambda) + \pi(k, i)\alpha_i \qquad k > 0,\ i = 1, \ldots, B. \tag{6.4}$$

$$\pi(0, 0i)(\beta + \lambda) = \pi(0, i)\alpha_i \qquad i = 1, \ldots, B. \tag{6.5}$$

The Markov chain above can then be solved numerically and used to obtain the exact system performance metrics.

6.3.2 Analytical Approximations

The previous Markov chain can be numerically solved, however it could be computationally intensive for large number of channels and large queues. We thus propose to use two different approximations based on the symmetric structure of this Markov chain to obtain analytical results for this model which could be used, for instance, in real-time CR network scheduling and configuration algorithms. Let us first define Q_i as the steady-state probability of operating over channel i which implies either the channel being used now is i (being in a row with an i index) or i was the last channel before a recovery (being in a row with a $0i$ index). The Q_i values can be directly found from a Markov chain, illustrated in Figure 6.3, where each state represents a whole line in the bi-dimensional Markov chain and the transition probabilities (rates) are given by $\alpha_i's$ and q_{ij} values. Note that as β is unique and thus the same for all rows, in finding the steady-state probabilities of operation over channel type i, the recovery periods are neglected. In the first approximation approach, we neglect the dependence between the rows for each service rate to approximate the z-transform of the number of packets in the queue as

$$\pi_z(z) \approx \sum_{i=1}^{B} Q_i \pi_z(z,i), \tag{6.6}$$

where $\pi_z(z,i)$ is the z-transform of the distribution of the number of packets in the system for a queue with a fixed service capacity μ_i. That is, we approximate $\pi_z(z)$ as the weighted sum of independent Markov chains for different service rates, where the weights are the probabilities of having each service rate.

In this Markov chain, each service rate is itself modeled by a two-dimensional Markov chain with two lines : one line for the operating state and the second for the recovery period. The z-transform of each line cannot be obtained because it does not correspond to a complete probability distribution (i.e., the probabilities of a line do not sum to 1). To solve this problem, we introduce $P_{(k,i)}$ and $P_{(k,0i)}$, which are the conditional probabilities that the system is in states (k,i) and $(k,0i)$, respectively, given that the system is in an operating period or a recovery period, respectively. From the symmetric structure, it can be immediately seen that $\pi(k,i) = \frac{\beta}{\alpha_i+\beta}P_{(k,i)}$ and $\pi(k,0i) = \frac{\alpha_i}{\alpha_i+\beta}P_{(k,0i)}$.

Let $q_{ii} = 1$ (and thus $q_{ij} = 0$,$\forall j \neq i$) to decouple the service rates. We then obtain the following balance equations as a function of the conditional probabilities $P_{(k,i)}$ and $P_{(k,0i)}$:

$$P_{(0,i)}\beta(\alpha_i + \lambda) = P_{(1,i)}\beta\mu + P_{(0,0i)}\beta\alpha_i, \tag{6.7}$$

$$P_{(0,0i)}\alpha_i(\beta + \lambda) = P_{(0,i)}\beta\alpha_i, \tag{6.8}$$

$$(\lambda + \alpha_i + \mu)\beta P_{(k,i)} = \mu\beta P_{(k+1,i)} + \lambda\beta P_{(k-1,i)} + \beta\alpha_i P_{(k,0i)} \qquad k > 0, \tag{6.9}$$

and

$$(\lambda + \beta)\alpha_i P_{(k,0i)} = \lambda\alpha_i P_{(k-1,0i)} + \beta\alpha_i P_{(k,i)}, \qquad k > 0. \tag{6.10}$$

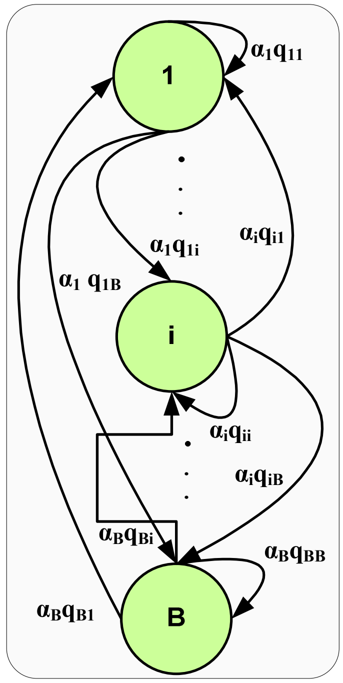

Figure 6.3 The steady-state probabilities of working on a channel of type i (Q_i) are the steady state probabilities of the Markov chain above.

Note that in all of the above equations, $(\alpha_i + \beta)$ is common in the denominator and was thus dropped. Because $P_{(k,i)}$ and $P_{(k,0i)}$ are complete probability distributions, we can then obtain their z-transforms $P_{z\,i}(z)$ and $P_{z\,0i}(z)$, respectively, by computing $\sum_{k=1}^{\infty} z^k$ on both sides of the last two equations. We can then find that $P_{z\,0i}(z)$ is equal to :

$$P_{z\,0i}(z) = \frac{\beta P_{z\,i}(z) - \beta P_{(0,i)} + (\lambda + \beta) P_{(0,0i)}}{(\lambda + \beta) - \lambda z} = \frac{\beta P_{z\,i}(z)}{(\lambda + \beta) - \lambda z} = P_{z\,i}(z) A_z(z). \tag{6.11}$$

$A_z(z)$ represents the z-transform of the number of arrivals during exponential recovery periods. $P_{z\,i}(z)$ can then be determined as :

$$P_{z\,i}(z) = \frac{\mu_i(1-\rho)(1-z)}{\mu_i(1-z) - \lambda z(1-z) - \alpha_i z(1 - \frac{\beta}{\lambda+\beta-\lambda z})}. \tag{6.12}$$

$P_{(0,i)}$ is equal to $1 - \rho_i = 1 - \lambda E[X_{b,i}]$ Azarfar *et al.* (2014d). We can now find $P_{(0,0i)}$ from (6.8), and from there, we can find that $\pi(0,i)$ and $\pi(0,0i)$. Finally, the $\mathbb{P}_0$ of the system can be found to equal :

$$\mathbb{P}_0 = \pi(0,i) + \pi(0,0i) = \frac{1-\rho_i}{1 + \frac{\lambda \alpha_i}{\beta(\alpha_i+\beta+\lambda)}}. \tag{6.13}$$

$\pi_z(z,i)$, the z-transform of the number of packets in an unconditioned queue with the rate

μ_i, can be found equal to :

$$\pi_z(z,i) = \frac{\beta}{\alpha_i+\beta} P_{z\,i}(z) + \frac{\alpha_i}{\alpha_i+\beta} P_{z\,0i}(z) = P_{z\,i}(z)(\frac{\beta+\alpha_i A_z(z)}{\alpha_i+\beta}). \tag{6.14}$$

The average number of packets in the system can be given by :

$$\begin{aligned}\pi_z^{'}(1,i) = \bar{N} &= P_{z\,i}{}'(1) + \frac{\alpha_i}{\alpha_i+\beta} A_z^{'}(1) \\ &= \frac{\rho_i}{1-\rho_i} + \frac{\alpha_i \lambda^2 E[R^2]}{2\mu_i(1-\rho_i)} + \frac{\alpha_i}{\alpha_i+\beta}\lambda E[R]. \end{aligned} \tag{6.15}$$

Combining those results with (6.25), the overall queue performance can be analytically approximated.

For the second approximation approach, we define the average service rate ($\bar{\mu}$) and the average failure rate ($\bar{\alpha}$) of the bi-dimensional queue as :

$$\bar{\mu} = \sum_{i=1}^{B} Q_i \mu_i, \tag{6.16}$$

$$\bar{\alpha} = \sum_{i=1}^{B} Q_i \alpha_i. \tag{6.17}$$

We then approximate the model with a system with homogeneous channels where the service rate is $\bar{\mu}$ and the operating periods are exponentially distributed with the parameter $\bar{\alpha}$. That is, the z-transform of the number of packets in the queue is approximated as :

$$\pi_z(z) \approx \pi_z(z, \bar{\mu}, \bar{\alpha}), \tag{6.18}$$

where $\pi_z(z,\bar{\mu},\bar{\alpha})$ is the z-transform of the distribution of the number of packets in the system for a queue with a fixed service capacity $\bar{\mu}$. $\pi_z(z,\bar{\mu},\bar{\alpha})$ can be found using the same approach as outlined previously to find $\pi_z(z,i)$.

6.4 Queue Model for General Interruptions

In this queue model, we simplify the system by assuming that all of the packets arriving during the recovery period actually arrive at the end of the recovery period, or equivalently, at the beginning of the next operating period. However, this model is a simplification and indeed a lower bound, as it neglects a small part of the waiting time for the packets that arrive during a recovery period. On the other hand, this assumption provides a lower bound analysis for generally distributed recovery periods. Exponential distributions for the operating periods and service time are still considered.

The evolution of the CR queue can be modeled with a continuous time Markov model, as illustrated in Figure 6.4. As the service rate may change between different operating periods, we use a two-dimensional Markov chain model where line i, $i = 1, \ldots, B$, corresponds to an operating period with channel i. State (k, i), $k = 0, 1, \ldots,$ and $i = 1, \ldots, B$ indicates that k packets are accumulated in the queue and that channel i with service rate μ_i and failure rate α_i is currently used. In each state of the Markov chain, in addition to transitions caused by the arrival and service processes, there are also transitions due to failure events. Those transitions depend on the channel selected after the recovery period and on the number of packet arrivals during this recovery period. For example, suppose that the current state is (k, i). If a failure occurs, the CR node performs a recovery and selects channel j, $j = 1, \ldots, B$ with probability q_{ij}. The next state will then be $\{(k, j), (k + 1, j), \ldots\}$, which depends on the number of arrivals (probability of n arrival, a_n, in R). The number of arrivals during the random recovery period R is denoted by the random variable A_r and its distribution is given by :

$$a_n = Pr(A_r = n) = \int_0^{\infty} \frac{e^{-\lambda t}(\lambda t)^n}{n!} f_R(t)dt, \tag{6.19}$$

where $f_R(.)$ is the probability density function of R.

In the proposed Markov chain, the *transition rate* from state (k, i) to state (l, j) can be expressed by :

$$p_{(k,i)(l,j)} = \begin{cases} \lambda + \alpha_i a_1 q_{ii} & i = j \text{ and } l = k + 1 \\ \mu_i & i = j \text{ , } l = k - 1 \text{ and } k > 0 \\ \alpha_i a_{(l-k)} q_{ij} & i \neq j \text{ and } l \geq k \text{ or } i = j \text{ and } l > k + 1 \\ 0 & \text{Otherwise.} \end{cases} \tag{6.20}$$

For the state $(0, i)$, the boundary steady-state balance equation can be written as :

$$(\lambda + \alpha_i)\pi(0, i) = \mu_i \pi(1, i) + a_0 \sum_{j=1}^{B} \alpha_j q_{ji} \pi(0, j). \tag{6.21}$$

For the states where $k > 0$, we have :

$$(\lambda + \alpha + \mu_i)\pi(k, i) = \mu_i \pi(k + 1, i) + \lambda \pi(k - 1, i) + \sum_{j=1}^{B} \sum_{n=0}^{k} a_n \alpha_j q_{ji} \pi(k - n, j), \tag{6.22}$$

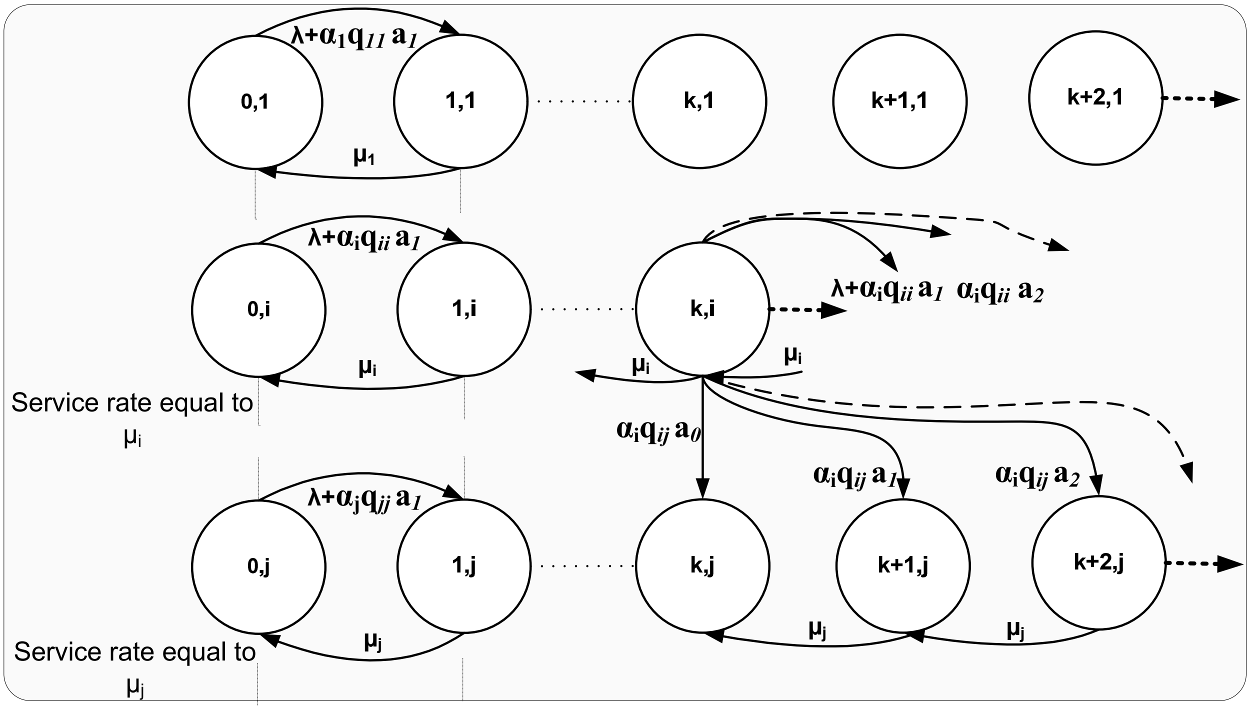

Figure 6.4 Approximate Markov chain for the queue with a variable service rate (M2).

and the last equation :

$$\sum_{i=1}^{B}\sum_{k=0}^{\infty}\pi(k,i) = 1. \tag{6.23}$$

Numerical techniques can then be used to solve these balance equations to find the steady-state probabilities $\pi(k,i)$ of this Markov chain. From this distribution, different performance parameters of interest can be obtained such as the queue occupancy, the system time and the waiting time. It is also straightforward to develop analytical approximations for this model using an approach similar to Section 6.3.2.

6.5 Simulation and Numerical Results

To investigate the accuracy of derived results, the first moment of the average number of packets in the system ($E[N]$) has been evaluated using the numerical solutions for the two proposed Markov models and the analytical approximations, and the results are compared with accurate Monte-Carlo simulations of the system. We investigated the CR performance for exponentially distributed operating periods and recovery periods with averages of 75 ms ($E[Y] = E[Y_i] = 1/\alpha_i = 75$ ms $\forall i$) and 25 ms ($E[R] = 1/\beta = 25$ ms), respectively. We also present results for a high system availability case where $E[Y] = E[Y_i] = 750$

ms ($\forall i$). We consider a system with $B = 6$ and rates $\mu_i = [0.27, 0.4, 0.47, 0.8, 0.87, 1]$ (packets/ms). The transition probabilities are independent of the original states and given by $q_{ij} = [0.3, 0.25, 0.15, 0.15, 0.1, 0.05]$ $\forall i$. In the figures, M2 refers to the lower-bound model presented in Section 6.4 and M1 to the accurate model for the exponential recovery period presented in Section 6.3, 'Num' refers to the Markov chain numerical solution, and 'An-Apx1' and 'An-Apx2' refer to the two approximate analytical solutions. Note that using simulation results, we find an appropriate buffer size to be able to truncate the Markov chains and solve them numerically.

As expected and illustrated in Figures 6.5, 6.6 and 6.7, the simulated $E[N]$ and $E[N^2]$ and their numerical evaluation using the first Markov chain are identical. As discussed, the second model provides a lower bound for the exact simulated performance. Regarding the analytical approximations, the first approximation is built on a weighted sum of B fixed-rate queues (Eq. (6.25)) and is therefore sensitive to instability in any one of the B queues. This approximation is therefore accurate as long as the weakest queue with the lowest service rate has a queue utilization, ρ_i, lower than one. For example, for $E[Y] = 75$ ms (Fig. 6.5), we can easily find that $\rho_1 = 1$ for $\lambda = 0.2025$. The first approximation thus starts to diverge when λ approaches 0.2025. The second approximation based on Eq. (6.25) underestimates $E[N]$, and the gap between the exact result and the approximation increases as a function of the arrival rate. This can be explained by considering that increasing the arrival rate makes the heterogeneity of the channels more important. This approximation thus underestimates the impact of the weakest queue on the overall performance by assuming homogeneous channels with a single average service rate. Figure 6.7 depicts the queue performance when availability periods are much longer than recovery periods (for this case $E[Y] = 30E[R]$). Longer availability periods imply that the system operates as a fixed-rate queue for long periods of time and the dependency between the different lines of the Markov chain thus becomes less important. The first approximation, which is based on this assumption, therefore provides a very tight bound for the exact performance of the queue.

The previous results have confirmed the exactitude of the Markov model for heterogeneous spectrum access and the first approximation tightness in the regime where all the channels are stable. In the following, we will use the accurate Markov model to gain some insights on the queue performance for an heterogeneous OSA network with two channel types. We assume a packet arrival rate of $\lambda = 0.12$ and a recovery period with an average length $E[R] = \frac{1}{\beta} = 10$. The parameters of the type 2 channel are an average availability period length of $E[Y_2] = 150$ and an average packet service time of $E[Y_2] = 7.5$. Figure 6.8 shows $E[N]$, the average number of packets in the system, as a function of q_1 ($q_2 = 1 - q_1$), the probability of selecting a type 1 channel after an interruption, for different parameters of the type 1 channel. $E[Y_1]$ varies

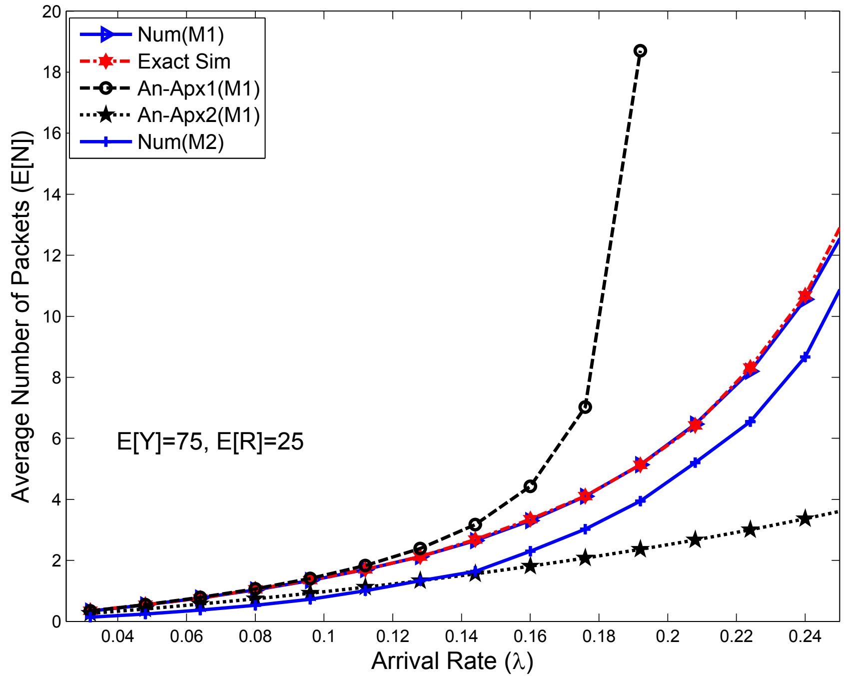

Figure 6.5 $E[N]$ versus the arrival rate λ.

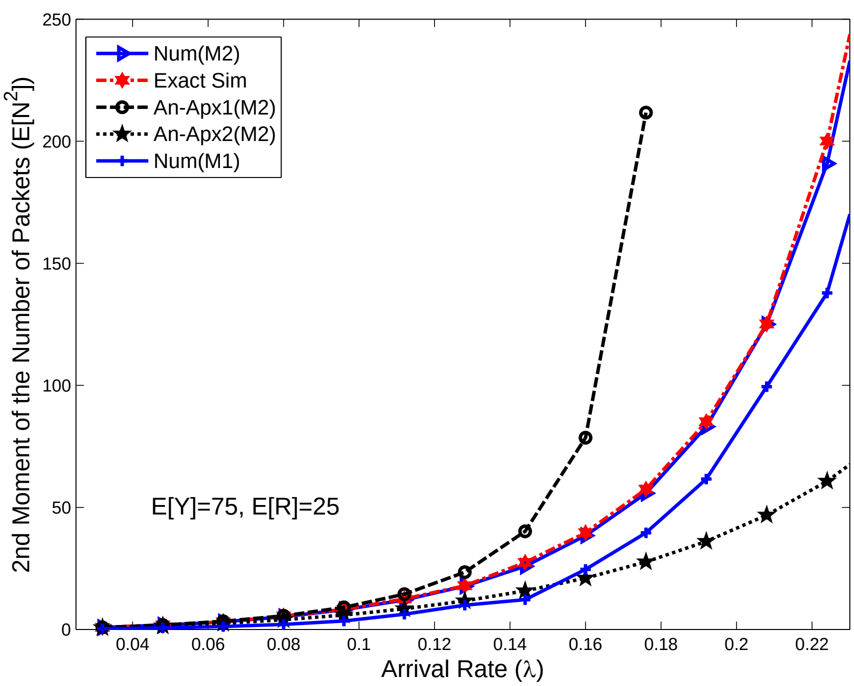

Figure 6.6 $E[N^2]$ versus the arrival rate λ.

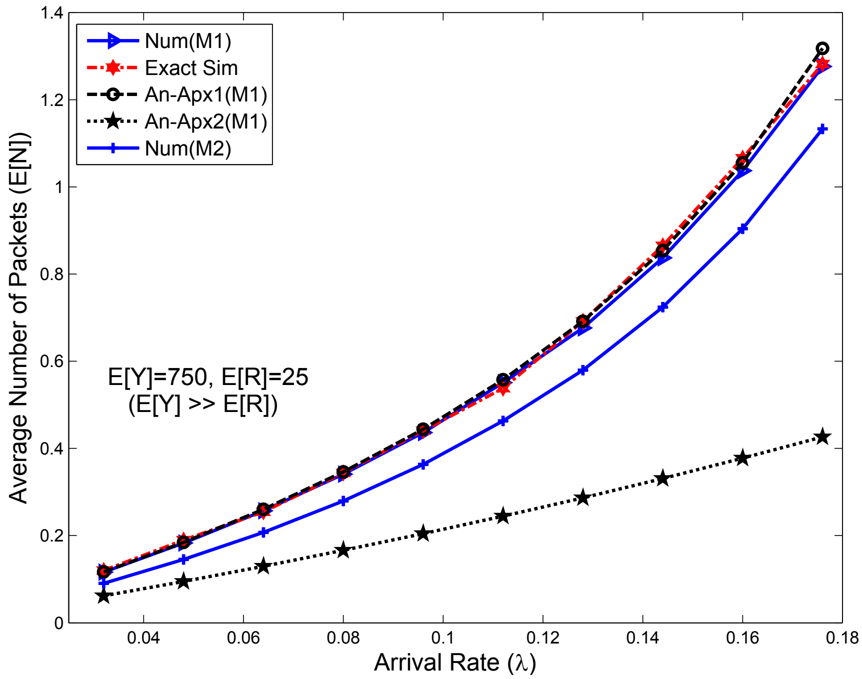

Figure 6.7 $E[N]$ when availability periods are much longer than recovery periods.

between 20 and 70 while $E[T_1] = 0.1E[Y_1]$ and therefore varies between 2 and 7. That is, the more reliable channels with longer availability periods have slower service rates.

First, it can be observed from this figure that $E[N]$ is a non-convex function of q_1 and that the lowest value of $E[N]$ is always achieved when only one channel type is used ($q_1 = 0$ or 1). We observed the same behavior for the minimal value of $E[N]$ for all system parameters that we studied. It is straightforward to see that this behavior will not be affected if we had different recovery distributions for each channel type, and if the average recovery time for a transition toward a channel type is longer than the time for staying on that channel type (which is normally the case since there is an additional switching time due to the transition between channels). We therefore conjecture that the optimal $E[N]$ in an heterogeneous OSA system, where the transition probabilities between channel sets are fixed, is achieved when the system always stays on the channel type with the minimum (single-row calculated) $E[N]$. However, due to the system complexity and non-convexity of $E[N]$, proving this conjecture remains an open problem.

Figure 6.9 shows the variance of N, which is related to the jitter and has an important impact on the buffer size design, as a function of the arrival rate for $E[T_2] = E[Y_2]/30$. Figures 6.8 and 6.9 demonstrate how the Markov chain analysis can be used for OSA network planning. For example, assume that the type 2 channels are the basic channels available to the OSA network and that the network designer needs to select the type 1 channel set. q_1

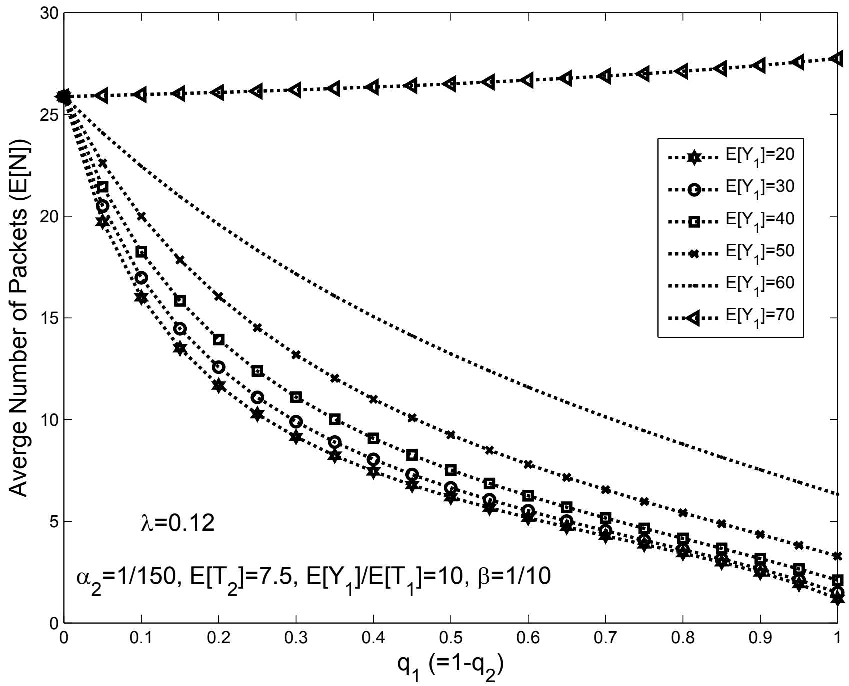

Figure 6.8 $E[N]$ versus q_1 for two types of channel ($B = 2$).

depends on this selection since it is a function of, without limitations, the average channel availability, the channel service rate, the number of available channels in the set and the number of users. It can be observed that depending on the value of q_1 for the different channel types, the optimal type 1 channel varies. For example, assuming that $q_1 = 0.2$ for $E[Y_1] = 20$, then in order to minimize $E[N]$, the type 1 channel set with $E[Y_1] = 40$ should be selected if for those channels we have $q_1 > 0.28$. Figure 6.9 also shows that there is multiple crossover points where the best channel set changes, further illustrating the non-convexity of the system performance and the importance of carefully selecting the channel sets as a function of the system parameters.

6.6 General Operating Periods and Service Times

Memoryless operating periods and service times were assumed in Sections 6.3 and 6.4 to develop Markov chain models. Although the memoryless model has been shown to be a good assumption Yin *et al.* (2012), we explore in this section the analysis and the performance of the CR queues with general distributions for the operating periods and service times.

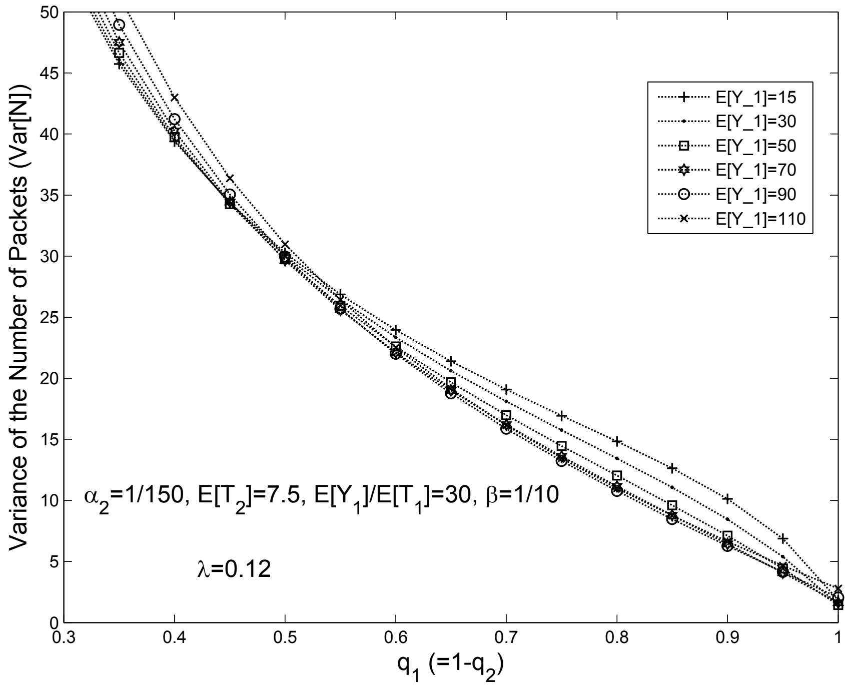

Figure 6.9 $Var[N]$ versus q_1 for two types of channel ($B = 2$).

6.6.1 Operating periods

The operating period distribution depends on several factors such as the remaining availability part of the selected channel during the recovery period. Since the system has no knowledge of how long this channel has been available so far, we should use renewal theory results Cox (1962); Ross (2006). Furthermore, the distribution of the remaining parts of the operating period will be different after each packet transmission, which further complicates the analysis. As discussed in Federgruen et Green (1986); Azarfar *et al.* (2012a), even for a fixed service rate queue with non-exponentially distributed operating periods, only analytical approximations can be obtained for the queue performance. An approximation for the heterogeneous system can be obtained by using the approximate analytical results for a queue with fixed service rate provided in Azarfar *et al.* (2012a) for each variable service rate queue into the two approximations (6.25) and (6.18). Results obtained by this approach are called 'Apx1-AnaApx' and 'Apx2-AnaApx' in the figure.

In Yin *et al.* (2012) it was discussed that one of the best models for the channels availability periods in cognitive radio networks is to assume a constant part in addition to an exponential tail. We therefore modeled the operating periods with a Pareto distribution with the same average as the exponential distribution used before ($E[Y]$ =75 ms) but a larger variance (shape parameter = 0.43, scale parameter = 18.3825 and threshold parameter = 42.75) . The constant part with this distribution is thus 42.75 ms. We also investigated the case where the CR system has constant operating periods (with the same average) to model a protocol where the system remains on a channel for a fixed given time and not until the moment where the channel becomes unavailable. Simulation and analytical results are presented in Fig. 6.10 for both cases as well as for the exponential distribution. We can observe that the impact of the higher moments of the operating periods on the average number of packets is small so that the queue performance has very close results for the different distributions. This can be explained by the fact that the packet service time is much smaller than the average operating periods. We can also observe that the average number of packets in the system for both Pareto and constant distributions is slightly lower than for the exponential distribution. This is due to the guaranteed constant part where no failure event can occur. We also see that the analytical approximations quality is similar for the general distributions as for the the exponential distribution. From those results, we can conclude that the performance results (either analytical or numerical with the Markov chain model) obtained with the memoryless model can provide acceptable approximations of the queue performance with generally distributed operating periods.

6.6.2 Service Time

For a generally distributed service time, we essentially have an M/G/1 queue with interruptions. For the average queue occupancy, from (Azarfar *et al.*, 2014d, Eq. 37) and for a queue with a fixed service rate we have :

$$E[N] = \lambda E[X_b] + \frac{\lambda^2 E[X_b^2]}{2(1-\lambda E[X_b])} + \frac{\lambda E[R^2]}{2(E[Y]+E[R])}, \tag{6.24}$$

where $E[X_b]$ is the average completion time Azarfar *et al.* (2012a) and is given by $E[T](1+\alpha E[R])$. $E[T]$ is the real service time of the packets (packet length divided by the queue service rate). Using (6.24) for each service rate, we can obtain the same analytical approximations with Eqs. (6.25) and (6.18) for the heterogeneous CR system. The two bounds will thus be given by $\sum_{i=1}^{B} Q_i E[N(\mu_i)]$ and $E[N(\bar{\mu})]$ (called 'Apx1-AnaApx' and 'Apx2-AnaApx' in the figures). We have performed simulations for two different distributions : Pareto distribution with a large variance (shape parameter = 0.43, scale parameter = 3.6765 and threshold parameter = 8.55) and constant packet size, both with the same average as for the exponentially distributed packet length in Section 6.5. As illustrated in Fig. 6.11, we can observe similar results for different packet length distributions and the two analytical approximations can still be used for performance analysis of a cognitive radio system with generally

Generalized Pareto distribution is used to generate Pareto instances.

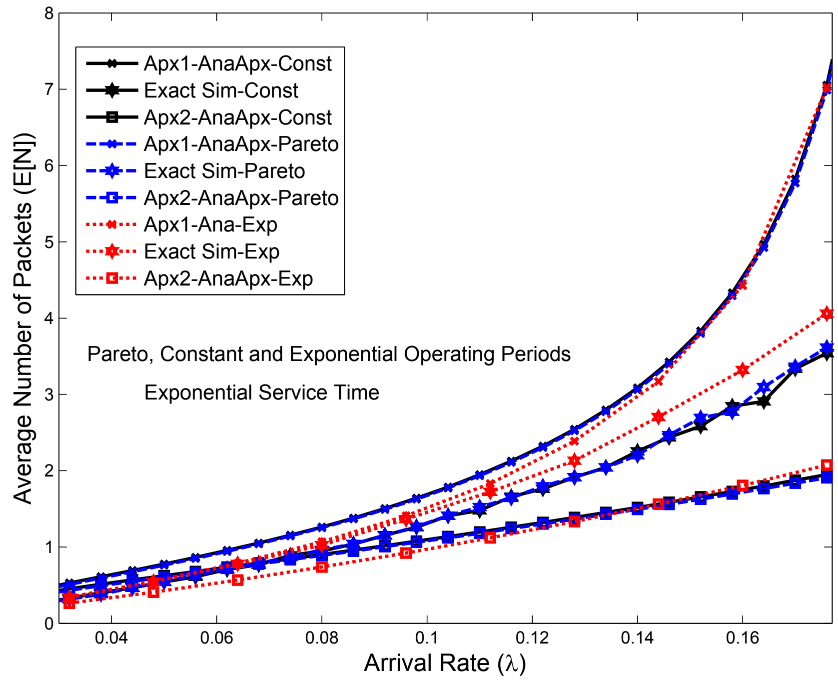

Figure 6.10 $E[N]$ when operating periods are distributed with Pareto and constant distribution ($E[Y] = 75$), compared to exponential distribution. All other parameters are the same as in Section 6.5.

distributed service times.

6.7 Conclusion

A queueing model with frequent interruptions and heterogeneous channels with variable service rates and failure rates was studied in paper for OSA in CR networks where the new channel after spectrum handover does not necessarily provide the same service rate and availability as the last channel. Modeling the queue as a two-dimensional Markov chain, we established numerical evaluations and analytical approximations for the general case in which recovery periods can have any distribution and for the specific case when they are exponentially distributed. Simulation results are presented to validate our analysis. We also investigated the performance for general distributions of the service time and operating periods and showed that in realistic scenarios, the performance is similar as for memoryless distribution.

This work constitutes a first step for the traffic level study of OSA networks. Future research includes the modeling of MAC protocols for heterogeneous OSA networks to study their performance using the analytical tools provided in this paper. Another interesting research area is to use dynamic transition probabilities between channel sets, for example as a function of the instantaneous number of packets in the system, to obtain a better performance than with the fixed transition probability scheme studied in this paper.

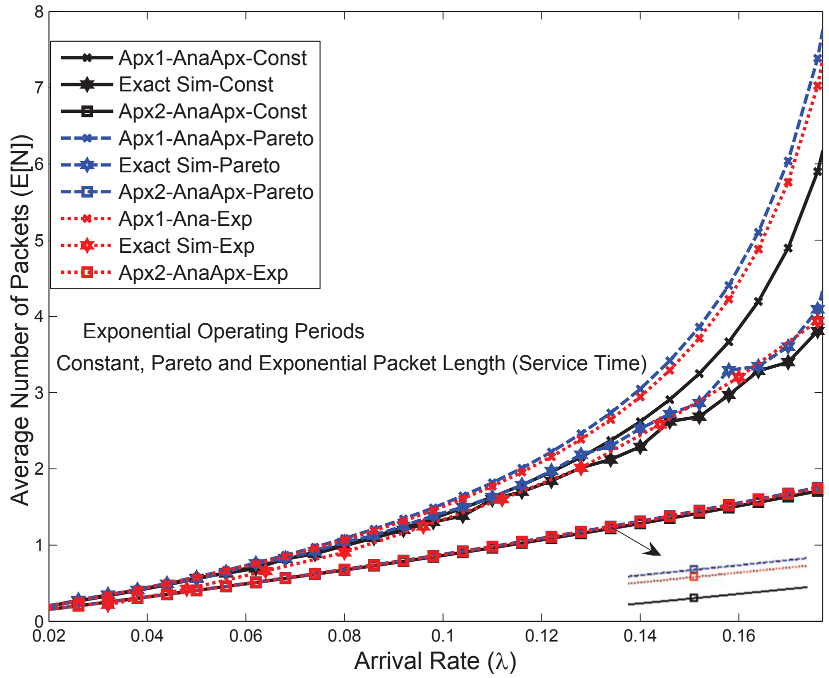

Figure 6.11 $E[N]$ when packet length is distributed with Pareto and constant distribution. The packet length average and all other parameters are the same as in Section 6.5.

6.8 Further discussions (Not a part of the paper)

6.8.1 z-Transform analysis for the approximation Markov chain

Due to space limit in the paper and considering the fact that the discussions are similar to the one provided for the accurate Markov chain, the z-transform analysis for the approximation Markov chain was dropped from the paper. It is provided here for completeness.

As discussed, in the first analytical approximation approach, we neglect the dependence between rows to approximate the z-transform of the number of packets in the queue as

$$\pi_z(z) \approx \sum_{i=1}^{N} Q_i \pi_z(z,i), \tag{6.25}$$

Here $\pi_z(z,i)$ is the z-transform of line i of the Markov chain illustrated in Figure 6.4 when $q_{ii} = 1$ (and thus $q_{ij} = 0$, $\forall j \neq i$), which removes the dependence between lines. This Markov chain has only one dimension and can be easily solved as follows. The steady-state balance equations are given by :

$$(\lambda + \alpha_i - \alpha_i a_0)\pi(0,i) = \mu_i \pi(1,i), \tag{6.26}$$

and

$$(\lambda + \alpha_i - \alpha_i a_0 + \mu_i)\pi(k,i) = \mu_i \pi(k+1,i) + \lambda\pi(k-1,i) + \alpha_i \sum_{n=0}^{k-1} a_{k-n}\pi(n,i). \quad (6.27)$$

After applying $\sum_{k=1}^{\infty} z_k$ on both sides and changing the order of $\sum$s (Kleinrock, 1975, P.135), we can write :

$$\begin{aligned}(\lambda + \alpha_i - \alpha_i a_0 + \mu_i)(\pi_z(z,i) - \pi(0,i)) = \\ \mu_i z^{-1}(\pi_z(z,i) - \pi(0,i) - z\pi(1,i)) \\ + \lambda z \pi_z(z,i) + \alpha_i \pi_z(z,i)(A_z(z) - a_0),\end{aligned} \quad (6.28)$$

where $A_z(z)$ is the z-transform of the distribution of the number of arrivals. We then have :

$$\pi_z(z,i) = \frac{\pi(0,i)\mu_i(1-z)}{\lambda z^2 + \alpha_i z A_z(z) - (\lambda + \alpha_i + \mu_i)z + \mu_i} \quad (6.29)$$

After letting $\pi_z(1,i) = 1$ and applying L'Hôpital's rule, $\mathbb{P}_0$, the probability that the system is empty can be found to be equal to :

$$\mathbb{P}_0 = \pi(0,i) = \frac{\mu_i - \lambda - \alpha_i A_z^{'}(1)}{\mu_i} = 1 - \lambda E[T_i](1 + \alpha_i E[R]) = 1 - \rho_i, \quad (6.30)$$

where $E[T_i] = 1/\mu_i$ is the average real service time (without taking the interruptions into account), $A_z^{'}(1) = A_z^{'}(z=1)$ is the average number of arrivals during a recovery period and is equal to $\lambda E[R]$, and $E[X_{b,i}] = E[T_i](1 + \alpha_i E[R])$ is the average *completion time* of the packets in the queue Azarfar *et al.* (2014d). The completion time is the total time that a packet spends in service including the real service time and interruptions that occur in-between Azarfar *et al.* (2014d). We also define $\rho_i = \lambda E[X_{b,i}]$ Azarfar *et al.* (2014d).

The z-transform of the steady-state probabilities $\pi_z(z,i)$, $i = 1, \ldots, B$ of the independent one-dimensional Markov chains can then be written as :

$$\pi_z(z,i) = \frac{\mu_i(1-\rho_i)(1-z)}{\mu_i(1-z) - \lambda z(1-z) - \alpha_i z(1 - A_z(z))}. \quad (6.31)$$

The average number of packets in the system can be found by $\pi_z^{'}(z=1,i)$, equal to :

$$\pi_z^{'}(1,i) = \frac{\rho_i}{1-\rho_i} + \frac{\alpha_i A_z^{"}(1)}{2\mu_i(1-\rho_i)}, \quad (6.32)$$

where $A_z^{"}(1)$ is the second moment of the number of arrivals during an interruption and is equal to $\lambda^2 E[R^2]$. Combining those results with (6.25), the overall queue performance can be analytically approximated.

6.8.2 Accuracy of the First Analytical Approximation

In this paper, two analytical approximations were discussed for the multi-row Markov chain, based on the accurate results which can be found for a fixed line. In this section, we want to discuss how accurate the first approximations may be based on the Jensen's inequality. As we have seen in Chapter 5, the function of the average queue occupancy versus $E[Y] = \frac{1}{\alpha}$ is a decreasing convex function similar to an exponential, as illustrated in Figure 6.12. In this figure, we considered only two different channel sets, as discussed in Section 6.5. The first analytical approximation is thus the line which connects two $E[N]$ values. If we assume that the function of $E[N]$ versus $E[Y]$ can be fitted with an exponential, Jensen inequality relates the analytical approximation and the accurate $E[N]$. We can thus see that if $E[Y_1]$ and $E[Y_2]$ are far from each other (and both are not too large), the analytical approximation (the line) will be far from the main curve. Further, when both $E[Y]$ values are large enough (on the tail of the curve), even if they are far from each other, the analytical approximation will be close to the curve and thus accurate. It therefore explains what has been observed in Figure 6.8 for $E[Y] = 70$ and $E[Y] = 150$, where although there is still 80 time unit difference between $E[Y_1]$ and $E[Y_2]$, the results are well matched. Note that in Figure 6.12, we assumed a fixed value of $E[T]$ while in Figure 6.8, $E[T_1]$ is also varying which may bring another source of difference between the numerical results and analytical approximation.

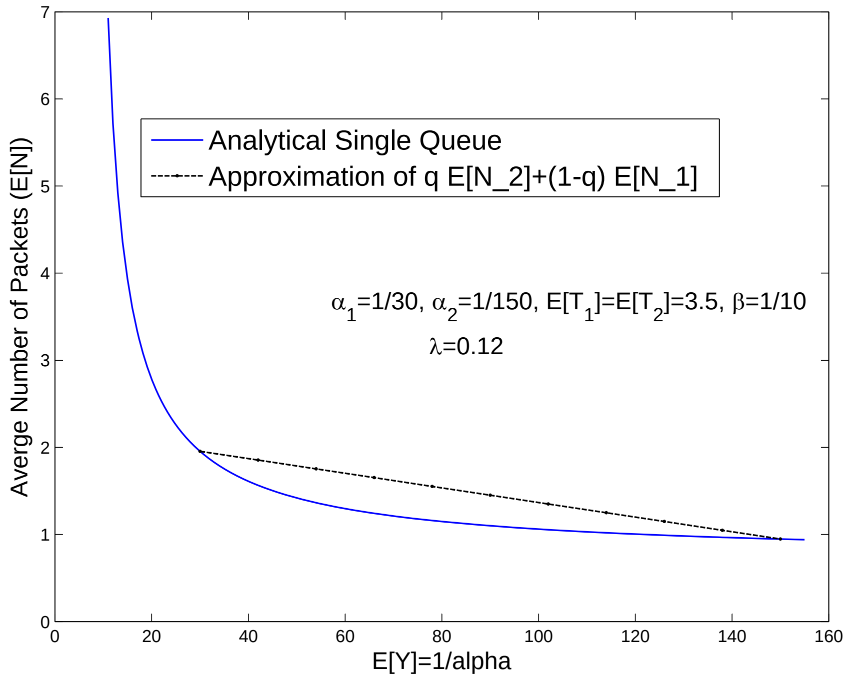

Figure 6.12 The accuracy of the first analytical approximation can be explained applying Jensen inequality to the curve of $E[N]$ versus $E[Y]$.

CHAPTER 7

ARTICLE 5 : DELAY ANALYSIS OF MULTICHANNEL OPPORTUNISTIC SPECTRUM ACCESS MAC PROTOCOLS

Arash Azarfar, Jean-François Frigon and Brunilde Sansò
Submitted to IEEE Transactions on Mobile Computing, March 2014
Department of Electrical Engineering
École Polytechnique de Montréal
C.P. 6079, succ. centre-ville, Montréal, QC, Canada, H3C 3A7
Email : {arash.azarfar, j-f.frigon, brunilde.sanso}@polymtl.ca

We provide a comprehensive delay and queueing analysis for two baseline medium access control protocols for multi-user cognitive radio networks and investigate the impact of different network parameters on the system performance. In addition to an accurate Markov chain, which follows the queue status of all users, several lower complexity queueing theory approximations are provided. Accuracy and performance of the proposed analytical approximations are verified with extensive simulations. It is observed that using an Aloha-type access to the control channel, a buffering MAC protocol, where in case of interruption the CR user waits for the primary user to vacate the channel before resuming the transmission, outperforms a switching MAC protocol, where the CR user vacates the channel in case of appearance of primary users and then compete again to gain access to a new channel. The reason is that the delay bottleneck for both protocols is the time required to successfully access the control channel, which occurs more frequently for the switching MAC protocol. We also propose a user clustering approach, where users are divided into clusters with a separate control channel per cluster, and observe that it can significantly improve the performance by reducing the competitions over control channel.

7.1 Introduction

Opportunistic spectrum access (OSA) communication models Zhao et Sadler (2007), implemented by cognitive radios (CR), offer both the capacity to decrease the communication infrastructure expenses by using the vacant portions of the spectrum and to improve the quality of wireless communications by permitting wireless CR nodes to switch to a bet-

---

AZARFAR, A., FRIGON, J.-F. and SANSO, B. (2014) Delay Analysis of Multichannel Opportunistic Spectrum Access MAC Protocols. Submitted to IEEE Transactions on Mobile Computing.

Node and *user* are used interchangeably.

ter channel when the channel quality is not satisfactory. These notable features have made cognitive-radio based wireless networks a technology of choice for incoming wireless technologies. Frequent spectrum handovers along with new requirements such as spectrum sensing and spectrum decision distinguish a cognitive radio network from its predecessors. New medium access control (MAC) protocols which take into account the inherent nature of cognitive radios are thus indispensable. For this aim, several medium access protocols have been proposed in the literature Zhao *et al.* (2007); Jia *et al.* (2008); Cordeiro et Challapali (2007). Different protocols may differ in details such as the existence of a common control channel, number of radios required for the nodes and sensing algorithm (please see Cormio et Chowdhury (2009); Pawelczak *et al.* (2009) and references therein for a detailed survey of CR MAC protocols). Since most of the proposed MAC schemes assume a dedicated control channel, and given that being equipped with multiple radios increases the hardware and complexity cost, we investigate in this paper a realistic cognitive radio network with a common control channel, distributed sensing appropriate for ad-hoc networks, and a single radio per node.

In the literature, most of the attention on the analysis of the different medium access protocols and their variations has been focused on the throughput analysis in the presence of saturated traffic (e.g., Pawelczak *et al.* (2009); Park *et al.* (2011); Zhao *et al.* (2007); Pawelczak *et al.* (2008); Jia *et al.* (2008); Cordeiro et Challapali (2007)), while little work has been done on delay and jitter analysis. Throughput analysis provides an upper bound for the performance of the network, which can be used for evaluation and comparison purposes of different protocols. However, delay analysis and packet level performance evaluation are required to investigate how a protocol, along with its parameter settings, behaves for delay-sensitive applications, such as multimedia communications Wang *et al.* (2011); Cisco (2013).

In the few papers available on delay analysis (e.g., Rashid *et al.* (2007); Wang *et al.* (2012b,a); Laourine *et al.* (2010); Kim (2012)), detailed system parameters are not considered or the CR network under consideration is simple, such that full insights into the impact of different parameters on the network performance have not been provided. The target has also been specific networks ; e.g., sensor networks, with specific requirements Bicen *et al.* (2012), such that the results can not be easily extended to general models. In Azarfar *et al.* (2012a, 2014d), we proposed a general M/G/1 queueing model Kleinrock (1975) for a cognitive radio link where the server is subject to interruptions Takagi (1991). The channel (i.e., the server of the queue) is subject to interruptions because the CR node has to suspend its communication session for the period of spectrum handover (switching) or primary users' activity (buffering). The proposed models in those papers were for a single node with continuous-time operating and interruption periods. However, those models didn't address the complex problem of multi-node operation in a multi-channel environment.

Our objective in this paper is thus to fill this MAC protocols analysis gap by providing a comprehensive delay analysis for two baseline medium access control protocols discussed in Pawelczak *et al.* (2009); Park *et al.* (2011). Note that the models described in those papers are not detailed MAC models for practical purposes. The intent is rather to have models with enough details, yet general enough, to use as baseline reference models for families of MAC protocols and investigate the effect of different parameters on the throughput of a cognitive radio network. In this paper, we have the same objective for the delay analysis and thus only relax the saturated traffic assumption and keep the same other assumptions and modeling details to analyze the average delay. In the first part, a buffering MAC model Pawelczak *et al.* (2009); Park *et al.* (2011), in which a node stays on the channel in case of interruptions until the transmission of the packet is finished, is investigated. In the second part, a switching MAC model Park *et al.* (2011), where the node leaves the interrupted channel and participates in a new reservation competition every time that an interruption occurs, is investigated.

The main contribution of this work is thus providing a comprehensive analytical delay analysis for a multichannel cognitive radio network with both buffering and switching recovery policies where the impact of different network parameters (number of users, number of channels, control channel access probability, channel availability, and arrival rate) on the performance is considered. The results can thus be used in network optimization and design problems. For both MAC models, we derive an exact Markov chain model which can be solved to obtain the system time distribution and moments. Since the number of states in the Markov chains grows exponentially with the number of nodes, we also propose lower complexity approximations based on discrete-time M/G/1 queueing theory results. In both MAC models, the service time of a packet is composed of two parts : the time spent to compete with other users to reserve a channel (for the switching MAC model, the reservation period may occur multiple times during the service of a single packet), and the transmission time. We use Markov chain models, combined with Renewal theory results, to derive approximate distributions of both parts of the packet service time. The service time moments are then used to find the approximate average system time of packets in the OSA network for both MAC protocols. We also provide numerical results to validate the analysis and provide insight on the delay performance of the MAC protocols for different network parameters.

The paper proceeds as follows. In Section 7.2, the system model is presented and related parameters are introduced. The next two sections discuss the buffering model. In Section 7.3, an accurate Markov chain for queue occupancy is proposed from which different parameters of interest can be derived. As the queue occupancy Markov chain is not scalable, a service cycle analysis is provided in Section 7.4. Section 7.5 discusses the switching model where the same approach is followed : first an accurate Markov chain for queue occupancy is proposed and then a delay cycle analysis and related approximations are provided. Numerical and simulation results for both buffering and switching models are discussed in Section 7.6. Finally Section 7.7 concludes the paper with some remarks on future research direction. The notation used in this paper are listed in Table 7.1.

7.2 System Model

In this section we provide a summary of the two generic multichannel OSA MAC protocols for which we derive the delay analysis under unsaturated traffic in later sections. More details on those protocols, as well as their throughput performance under saturated traffic, can be found in Pawelczak *et al.* (2009); Park *et al.* (2011). It is assumed that there are N nodes in an ad-hoc cognitive radio network and M OSA channels where one channel is the dedicated control channel and the remaining $M_C = M - 1$ channels are used for data transmission. Without loss of generality and for notation simplicity, we assume that the N nodes have intended receivers that are not part of those nodes. A maximum of $s_{max} = \min(N, M_c)$ links might thus simultaneously exist.

As illustrated in Fig. 7.1, operations in the OSA network is time-slotted. Primary and secondary users are fully synchronized. In the beginning of each timeslot, there is a short quiet period during which the whole cognitive network stops operation to sense the channels. Primary users (PU) activity in any channel, including the control channel, is modeled by a Bernoulli random variable : in each timeslot and independent of the other timeslots and channels, a channel is unavailable for CR users (occupied by primary users or due to false alarms) with a probability p_c. We also use a packet capture model whereby a transmission during an available timeslot is successful, without considering collisions and interference, with probability η for the data channels and probability η_C for the control channel. Channel availability and packet captures are assumed to be independent.

A geometric (Bernoulli) arrival process is assumed for all nodes with a probability of packet arrival in a timeslot λ. The packet length L is given in the number of timeslots required to transmit the packet and has a geometric distribution with parameter q.

A node is in the busy state when it has a reserved data channel for transmission, otherwise the node is in the idle state (with an empty or non-empty packet queue). After the sensing period, if the control channel is available, the g idle nodes with a non-empty packet queue attempt to reserve a channel using an Aloha-type competition over the control channel. That is, each of the g nodes will transmit in the timeslot a reservation request with a probability p over the control channel. The competition is successful if only one reservation request is transmitted and it is correctly received. The probability of success in competition is given

Table 7.1 Notations

Notation	Description
N	Number of CR nodes
M	Total number of channels
M_C	Number of data channels ($= M - 1$)
λ	Poisson packet arrival rate
X	Geom/G/1 service time
X_R	Reservation time
X_T	Transmission time (buffering)
X_U	Interrupted transmission time (switching)
L	Packet length
L_e	Enlarged packet length (switching)
g	State variable, number of competitors
k	State variable
p_c	Probability of unavailability due to PUs
χ	Control channel availability per timeslot
η	Packet capture probability
ψ	Data channel availability $(1 - p_c) * \eta$
s_{max}	Maximum No. of links equal to $\min(N, M_C)$
$H_{s_{max}}$	Probability of k being equal to s_{max}
$P_s(g)$	Probability of success in Aloha-type competition
$P_s(k,g)$	Probability of success in state(k,g) (7.6)
$P_s^m(k,g)$	Probability of success in state(k,g) for a marked node
p	Probability of channel access attempt in Aloha-type competition
q	Packet length
Qs_{max}	Maximum Queue Size
$T_k^{(j)}$	Probability of j transmission among k
$S_g^{(j)}$	Probability of j successful reservation
$Y_a^{b,c}$	Defined in (7.8)
$\pi_{(k,g)}$	Steady-state probability of state (k,g)
$\pi_{(k,g)}^N$	Updated steady-state probabilities, eliminating states with $g = 0$
A_u	Inter-arrival time of PUs (switching model)

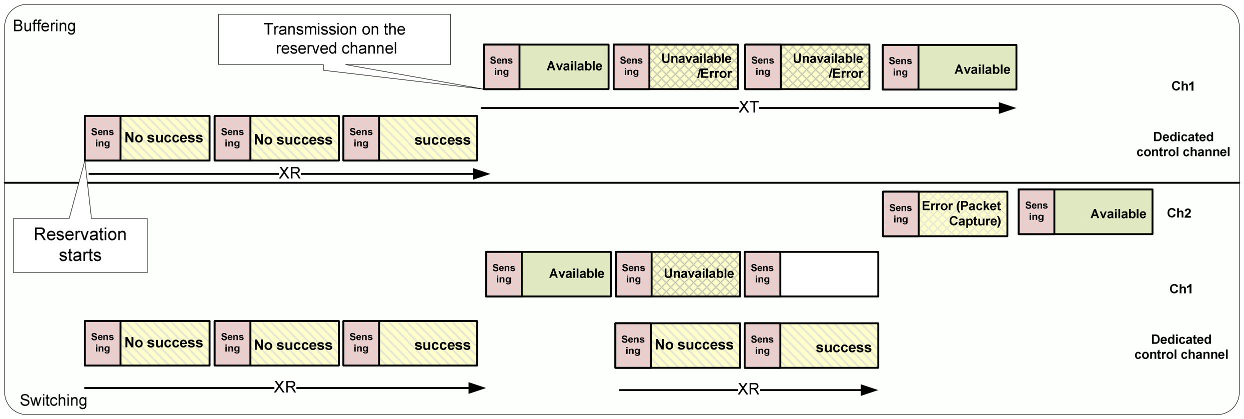

Figure 7.1 An example of the reservation and transmission processes. In the buffering model, a node reserves channel $Ch1$ in three timeslots ($X_R = 3$). The transmission of a packet with the length 2 (timeslots) takes 4 timeslots to finish ($X_T = 4$). The service time for this packet is thus $X = 7$ timeslots. In the switching model, after a channel unavailability in the second timeslot of the transmission on Ch1, the node participates in a new competition and reserve Ch2 after two timeslots. We thus have $X = 8$.

by Pawelczak *et al.* (2009) :

$$P_s(g) = gp(1-p)^{g-1}\chi, \tag{7.1}$$

where $\chi = (1-p_c)\eta_C$ is the probability of a successful transmission on the control channel without considering collisions. The probability of success for a marked node among g nodes is $\frac{P_s(g)}{g}$. It is assumed that the list of channels not being used by CR nodes is given. So, a success in competition in this timeslot is followed by reserving the channel which is on the top of the list by the competition winner, to be used from the beginning of the next timeslot. An exception is when all channels are occupied in this timeslot. If a channel is released at the end of this timeslot, it will be assigned to the successful node in competition, otherwise the successful node in competition is unsuccessful in reserving a channel and returns for a new competition in the next timeslot. It is therefore important to distinguish the event of *success in competition*, which depends solely on the number of competing users, from the event of *success in reservation*, which depends also on the status of channels. In case that a node is successful in reserving a channel, it starts transmitting one packet in the reserved channel from the beginning of the next timeslot. Other nodes can not make any interference on the reserved channel, but fading, sensing errors and PU activities may interrupt the transmission.

Two MAC protocols are considered in this paper. A packet transmission cycle example for each protocol is illustrated in Fig. 7.1. In the buffering MAC protocol Pawelczak *et al.* (2009), a node stays on its reserved channel, even when it becomes occupied by PUs, until the packet is entirely transmitted. In each timeslot, a successful transmission therefore occurs with probability $\psi = (1-p_c)\eta$. For the buffering MAC protocol, the (enlarged) packet service time X therefore consists of a reservation period of length X_R followed by a transmission time X_T which consists of successful transmission, unsuccessful transmission and unavailable timeslots, until the entire packet is transmitted.

In the switching MAC protocol Park *et al.* (2011), a node that senses it channel occupied by PUs leaves the channel and returns to the idle state to participate in the competition to reserve a new channel in this timeslot. For the switching MAC protocol, the (enlarged) packet service time X therefore consists of several reservation periods and the successful transmission and unsuccessful transmission timeslots required to transmit the entire packet.

For both MAC protocols, once the packet transmission is terminated, the node releases the channel and returns to the idle state. It will enter the competition at the next timeslot if it has another packet in its queue. A service-resume transmission model is assumed for both MAC protocols where the remaining part of the packet is transmitted after each interruption. Note that due to the packet length geometric distribution model, the delay analysis is also the same for a service-repeat model where the entire packet must be transmitted after each interruption.

7.3 Buffering MAC Protocol : Queue Occupancy Markov Chain Analysis

The discrete and memoryless nature of the system model and MAC protocols described in Section 7.2 enables us to propose a general discrete-time Markov chain (MC) which tracks all arrivals, departures and interruption events for all nodes. The different queue performance metrics can then be directly computed using the steady-state MC solution. The state variable is given in the form $(\vec{n}, \vec{b}) = (n_1, n_2, \dots, n_N, b_1, b_2, \dots, b_N)$ where n_i ($n_i = 0, 1, \dots$) is the current number of packets in the system (waiting in the queue or being transmitted) of node i and b_i ($b_i = 0, 1$) is a status flag indicating whether node i has a reserved channel ($b_i = 1$) or is idle ($b_i = 0$). For each state A, we also define the variables $k_A = \sum I_{(b_i=1, n_i>0)}$ and $g_A = \sum I_{(b_i=0, n_i>0)}$, where I is an indication function with $I_{true} = 1$, $I_{false} = 0$, which are, respectively, the total number of busy nodes and the total number of idle nodes with a non-empty queue. An MC example with $N = 1$ is illustrated in Fig. 7.2.

Because of the geometric arrival process and since at maximum one packet can be served in one timeslot over each channel, the queue status changes at most by one packet in two consecutive slots. That is $n_i(t) - 1 \leq n_i(t+1) \leq n_i(t) + 1$, where $n_i(t)$ is the node i queue occupancy at time t. However, as a function of the status flag, the following refinements can be made :

1. If $n_i = 0$ then $b_i = 0$, and only an arrival event can occur, in which case $n_i(t+1) = 1$,

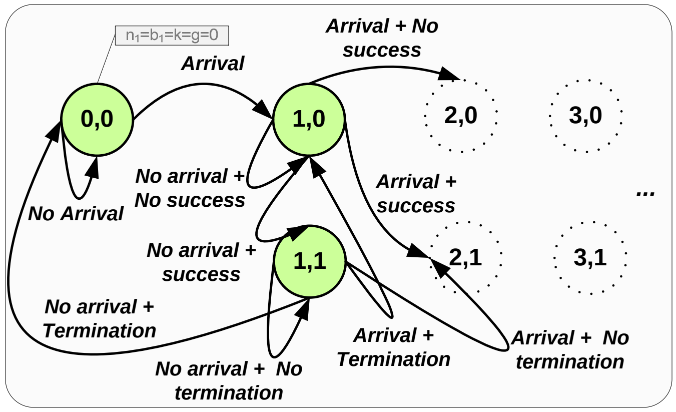

Figure 7.2 An example of the proposed queue occupancy Markov chain for $N = 1$ and $M_C > 0$.

otherwise $n_i(t+1) = 0$. Furthermore, the next status state is also $b_i = 0$, irrespectively of an arrival event occurrence.

2. If $n_i > 0$ and $b_i = 0$, two events of reservation success and an arrival can occur at the same timeslot and the events are independent. We have $n_i(t) \leq n_i(t+1) \leq n_i(t) + 1$ depending on arrival or no arrival. Reservation success only changes the value of b_i to 1.

3. If $n_i > 0$ and $b_i = 1$, two events of packet transmission termination and an arrival can occur at the same timeslot and the events are independent. We have $n_i(t) - 1 \leq n_i(t+1) \leq n_i(t) + 1$. The case $n_i(t+1) = n_i(t) + 1$ occurs with an arrival and no transmission termination, and the case $n_i(t+1) = n_i(t) - 1$ occurs with no arrival and transmission termination. Meanwhile, a node will have the same number of packets in the queue in the next timeslot for two events : transmission termination and no arrival, or no transmission termination and no arrival. A transmission termination results in changing the value of b_i to 0, irrespectively of an arrival event occurrence.

Furthermore, all busy nodes and idle nodes with an empty queue operate independently, so the joint probability of different events can easily be found by multiplication.

For two states $A = (\vec{n}, \vec{b}) = (n_1, \ldots, n_N, b_1, \ldots, b_N)$ and $B = (\vec{m}, \vec{c}) = (m_1, \ldots, m_N, c_1, \ldots, c_N)$, $\mathbb{IS}$, the set of impossible transitions, can then be defined as follows :

$$\mathbb{IS} = \Bigg\{ (A,B) = ((\vec{n},\vec{b}),(\vec{m},\vec{c}))|$$

$$\exists n_i, m_i\ (n_i = 0\ \&\ b_i = 1)\ \|(n_i = 0\&c_i = 1)\ \|$$

$$(m_i = 0\ \&\ c_i = 1)\ \|(Num_{s,AB} > 1)\ \|(m_i \notin (n_i - 1){:}(n_i + 1))\|$$

$$\Big(\{k_A = s_{max}\}\&\{Num_{(s,AB)} > 0\}\&\{Num_{(t,AB)} = 0\} \Big) \Bigg\}, \tag{7.2}$$

where $Num_{(s,AB)} = |find(\vec{c} - \vec{b} == 1)|$ ($|find(\vec{x} - \vec{y} == 1)|$ is the number of positions for which the difference between the elements of $\vec{x}$ and $\vec{y}$ is 1) is the number of nodes which started transmitting in the new timeslot from state A to B and $Num_{(t,AB)} = |find(\vec{b} - \vec{c} == 1)|$ is the number of busy nodes in state A whose transmission is finished in state B.

The transition probabilities of going from state A to a new state B can then be computed using Alg. (1). For $(A, B) \in \mathbb{IS}$, the transition probability is zero. In Alg. (1), $\mathbb{S}_g$ is the set of nodes participating in competition in state A whose size is given by $|\mathbb{S}_g|$. If $\mathbb{S}_g$ is empty, the last part of the algorithm is not run. $AllBusyNoTer = 1$ represents the case where all channels are busy and no termination occurs.

Algorithm 1 The algorithm to find the transition probabilities $P_{A=(\vec{n},\vec{b})->B=(\vec{m},\vec{c})}$.

```
if (k_A == s_max & Num_(t,AB) == 0) then
    AllBusyNoTer = 1
end if
for i = 1 : N do
    if (n_i == 0) then
        P_i = λ^(m_i − n_i) (1 − λ)^(1−(m_i − n_i))
    else
        if b_i == 1 then
            if (c_i == 1) then
                P_i = (1 − qψ)λ^(m_i − n_i) (1 − λ)^(1−(m_i − n_i))
            else
                P_i = (qψ)λ^(1−(n_i − m_i)) (1 − λ)^(n_i − m_i)
            end if
        else
            Put i in S_g
            P_i = λ^(m_i − n_i) (1 − λ)^(1−(m_i − n_i))
        end if
    end if
end for
***** Handling Competitions ***** (S_g ≠ ∅)
Find j s.t. (c_(S_g[j]) == 1)
if j > 0 then
    P_(S_g[j]) = P_(S_g[j]) * P_s(|S_g|)/|S_g|
else
    if (AllBusyNoTer == 0) then
        P_(S_g[1]) = P_(S_g[1]) * [1 − P_s(|S_g|)]
    end if
end if
FindProb(A,B)=∏_i^N P_i
```

This exact Markov chain model is general and can address several system model extensions, such as non-homogeneous nodes with different arrival rates or different packet length. However, as there is no buffer limit, the number of states is infinite and grows exponentially with the number of nodes. To be able to solve the Markov chain numerically a buffer limit Qs_{max} should be set and the truncated Markov chain is analyzed instead. If the transition probabilities are written for a truncated version of the Markov chain with the buffer size Qs_{max}, the states with $\exists i$ s.t. $n_i = Qs_{max}$ should be treated separately because for cases $b_i = 0$, or $b_i = 1$ and no termination, the queue occupancy in the next state with arrival or without any arrival (with probability equal to one) will still be $m_i = Qs_{max}$.

Even for the truncated MC, the queue occupancy MC model suffers from exponential growth of the number states and is thus not scalable. For example, for a buffer limit of Qs_{max}, the number of states will be $[2(1+Qs_{max})]^N$ (e.g., 484 states for $N = 2$ and $Qs_{max} = 10$). In the next section, the homogeneity of the nodes is taken into account to propose an approximate but scalable combined Markov chain which will be used for a service cycle analysis.

7.4 Buffering MAC Protocol : Service Cycle Analysis

Since a geometric arrival process is assumed, we have for any node a discrete-time M/G/1 queue with geometric arrivals called Geom/G/1. Continuous-time M/G/1 and Geom/G/1 become equivalent when the timeslot is short with respect to the arrival rate. Furthermore, since the arrival process and packet length distribution are identical for all nodes, the queues are homogeneous. To solve a tagged node queue, we should therefore only find the service time $X = X_R + X_T$ for the buffering MAC protocol. X_R and X_T are independent, so the first and second moments of the service time X can easily be found from the moments of X_R and X_T. Discrete version of the Pollaczek—Khinchine formula Takagi (1993) can then be used to find the queue performance metrics of interest. The exact distribution of X_T and approximations for the distribution of X_R are derived in Section 7.4.1 and 7.4.2, respectively.

7.4.1 Transmission Time

For the buffering MAC protocol, the transmission time is started after the timeslot in which a channel has been reserved with success and lasts until the end of the packet transmission (see Fig. 7.1). Since the node is using a reserved channel, there is no interference from other CR nodes and in each timeslot the probability of a successful transmission is given by $\psi = (1 - p_c)\eta$ (see Section 7.2). For a packet with length $L = n$ timeslots, n successful transmissions are required to complete the packet transmission. For $X_T = k$ timeslots, the last slot should necessarily be a successful transmission and among the remaining $k-1$ timeslots, exactly $n-1$ of them should be successful transmissions. Therefore, conditioned on the packet length $L=n$, the transmission time of the packet X_T has a negative binomial distribution given by :

$$Pr(X_T = k|L = n) = \binom{k-1}{n-1}\psi^n(1-\psi)^{k-n},\ k = n, n+1, \ldots \tag{7.3}$$

The packet length L has a geometric distribution such that the unconditional distribution of X_T is given by :

$$Pr(X_T = k) = \sum_n Pr(X_T = k|L = n)f_L(n) = \sum_{n=1}^{\infty}\binom{k-1}{n-1}\psi^n(1-\psi)^{k-n}(1-q)^{n-1}q. \tag{7.4}$$

The first and second moment of X_T can then be found as :

$$E[X_T] = \frac{1}{q\psi},\ E[X_T^2] = \frac{2-q\psi}{(q\psi)^2}. \tag{7.5}$$

7.4.2 Reservation Periods

At the beginning of every timeslot, nodes can be divided into three groups : busy nodes with a reserved channel, idle nodes with an empty queue and idle nodes with a non-empty queue. The probability of reservation success at the end of the timeslot depends on the number of nodes in the last group, except for the case where all channels are occupied. Due to packet arrival and packet transmission termination events, the number of idle nodes with a non-empty queue changes from one timeslot to the next one such that the probability of reservation success in a timeslot is time-varying. We will proceed as follows to find the reservation period distribution. First, we will propose a combined Markov chain to find the distribution and transition probabilities of the number of nodes competing in the reservation procedure. We will then find the reservation period distribution by first conditioning it on the number of competing nodes in the first reservation timeslot for the tagged node, and then using the steady-state probabilities to obtain the unconditioned distribution.

Combined Markov Chain

Since the sum of nodes at every timeslot is fixed to N, the state variable of the proposed combined Markov chain is given by a 2-tuple (k, g) where k is the number of busy nodes and g is the number of idle nodes with a non-empty queue. The number of idle nodes with an empty queue is then given by $N - k - g$. We have the following facts about the system :

– Every idle node with an empty queue may independently become a node with a non-empty queue in the next timeslot with the probability $P_a = \lambda$, which is the probability of one arrival during one timeslot.

– Among the idle nodes with a non-empty queue, at most one node may become busy with the probability of reservation success $P_s(k, g)$ and all others stay in the same group. $P_s(k, g)$ is the probability of getting a channel when g nodes participate in competition and k channels are already reserved. It is given by :

$$P_s(k,g) = \begin{cases} P_s(g) & k < s_{max} \\ P_s(g)(1 - T_k^{(0)}) & k = s_{max} \end{cases} \tag{7.6}$$

where $T_k^{(0)}$ is the probability of no transmission termination among the k ongoing communication sessions.

– Every busy node may independently finish in a timeslot its packet transmission with probability $q\psi$ and then, with probability P_0, the node joins the idle nodes with an empty queue. $1 - P_0$ converges to the steady-state queue occupancy of a node queue and is the same for all busy nodes because nodes are homogeneous and have a geometric memoryless packet length.

The number of participants in the competition in the next slot compared to the previous slot may thus decrease at most by one, but may increase to any larger value up to N depending on the number of idle nodes with an empty queue with an arrival and the number of busy nodes that finish their transmission and have a non-empty queue.

Using these facts and the results in (Pawelczak *et al.*, 2009, Eq. (16)), the transition probabilities of the combined Markov chain can be given by :

$$P_{(k,g)->(z,h)} = \begin{cases} 0 & \text{if } z > k+1, \forall g, h, \\ 0 & \text{if } \forall k, z, h < g-1, \\ T_k^{(0)} S_g^{(1)} Y_{(h-g+1)}^{(k-z+1),(N-k-g)} & \text{if } z = k+1, h \geq g-1, \\ T_k^{(k-z)} S_g^{(0)} Y_{(h-g)}^{(k-z),(N-k-g)} + & \\ T_k^{(k-z+1)} S_g^{(1)} Y_{(h-g+1)}^{(k-z+1),(N-k-g)} & 0 < z \leq k, k+z \neq 2s_{max}, h \geq g-1, \\ T_k^{(k-z+1)} S_g^{(1)} Y_{(h-g+1)}^{(k-z+1),(N-k-g)} + & \\ T_k^{(k-z)} Y_{(h-g)}^{(k-z),(N-k-g)} & \text{if } z = k = s_{max}, h \geq g-1, \\ T_k^{(k)} S_g^{(0)} Y_{(h-g)}^{(k-z),(N-k-g)} & \text{if } z = 0, h \geq g-1. \end{cases} \tag{7.7}$$

$T_k^{(j)}$ and $S_g^{(j)}$ are respectively defined in (Pawelczak *et al.*, 2009, Eqs. (14),(15)) as the probability of j transmission terminations among k ongoing communication sessions and the probability of j success in competition among g competitors (note that $S_g^{(1)} = P_s(g)$).

$Y_{(a)}^{b,c}$ is the probability that a new nodes join the group of idle nodes with a non-empty queue given that there are b busy nodes that terminate their transmission and c idle nodes with an empty queue. This happens if i nodes out of the b busy nodes that terminate their transmission have a non-empty queue at the end of the timeslot and j nodes out of the c idle nodes with an empty queue have a packet arrival in the timeslot, where $i + j = a$. $Y_{(a)}^{b,c}$ is thus given by :

$$Y_a^{b,c} = \sum_{(i,j)|i+j=a,\ 0 \leq i \leq b\ \&\ 0 \leq j \leq c} \binom{b}{i} (1-P_0)^i (P_0)^{b-i} \binom{c}{j} (P_a)^j (1-P_a)^{c-j}. \tag{7.8}$$

Assuming that P_0 is known, the steady-state probabilities $\pi_{k,g}$ $\forall (k,g)$ of the combined Markov chain can be found. We explain in Section 7.4.2 how P_0 can be found.

Reservation Period Distribution

We now derive the distribution of X_R for a tagged node belonging to the group of idle nodes with a non-empty queue. Let $X_R^{k,g}$ denote the reservation length given that the system is in the state $(k,g), g > 0$ at the beginning of the reservation period. The probability of a successful reservation for the marked node in the first timeslot, and thus have $X_R^{k,g} = 1$ is equal to $P_s^m(k,g) = \frac{P_s(k,g)}{g}$. However, if the reservation is unsuccessful, the system transits to a new state (z,h) with the probability $P_{(k,g)->(z,h)}|$(No-Success). In this state, the probability of successful reservation and thus have $X_R^{k,g} = 2$ is now $P_s^m(z,h)$. Otherwise, the system transits to a new state and the process repeats. In general, $\Pr(X_R^{k,g} = i)$ can be found in a recursive manner from the probability of transition to other states (z,h) and $\Pr(X_R^{z,h} = i-1)$ from other states, given that the node was not successful in reservation.

It is important to note that for this analysis, the transition probabilities of the combined MC discussed in the previous section can not be directly used because it is implicitly assumed that the marked node was not successful. We therefore need a different MC with transition probabilities denoted by Q conditioned on the fact that at each transition, the marked node was not successful, and that if the reservation competition was a success, it was one of the other $g-1$ competitors who was successful. The transition probabilities Q are given by (7.7) where $P_s(g)$ given in (7.1) is replaced by $(g-1)p(1-p)^{g-1}\psi$.

The distribution of $X_R^{k,g}$ is then given by :

$$\Pr(X_R^{k,g} = i) = \begin{cases} P_s^m(k,g) & i = 1 \\ (1 - P_s^m(k,g)) \sum_{(z,h)} Q_{(k,g)->(z,h)} P_s^m(z,h) & i = 2 \\ (1 - P_s^m(k,g)) \sum_{(z,h)} Q_{(k,g)->(z,h)} (1 - P_s^m(z,h)) & \\ \qquad \sum_{(a,b)} Q_{(k,g)->(a,b)} P_s^m(a,b) & i = 3 \\ .... & \end{cases} \tag{7.9}$$

$X_R^{k,g}$ can then be unconditioned to obtain X_R as follows :

$$X_R = \sum_{(k,g),g>0} X_R^{k,g} \pi_{(k,g)}^N. \tag{7.10}$$

Since when we mark a node to find X_R, we implicitly assume that the node is idle with a non-empty queue, the states in which the number of competing users is zero $(g = 0)$ should therefore be eliminated to obtain the steady-states probabilities for the states $(k,g), g > 0$ as $\pi_{(k,g)}^N = \frac{\pi_{(k,g)}}{1-\sum_{(k,g)|g=0} \pi_{(k,g)}}$.

Solution Procedure

To find the combined MC transition and steady-state probabilities, P_0, the steady-state probability of a node's queue being empty, is required. P_0 can be computed from the Geom/G/1 queue relation $(1-P_0) = \lambda E[X] = \lambda E[X_R + X_T]$. However, $E[X_R]$ depends on the combined MC solution. We therefore proposed an iterative solution where an initial value is set for P_0. The combined MC chain can then be solved and $E[X_R]$ computed. A new value of P_0 is then obtained and used for the next iteration. The iteration continues until the value of P_0

converges. To compute the initial value of P_0, we propose to use the lower bound case for X_R where only the tagged node participates in the competition. The distribution of X_R is then geometric and $E[X_R] = (p\chi)^{-1}$.

Note that the number of competing nodes in one state is not independent of the other states. However by using the steady-state probabilities in (7.10) combined with the iterative solution procedure, this dependency is ignored. Furthermore, the different functions involve in this system are not necessarily convex. The combined-MC approach is therefore inherently an approximate solution. For example, we have observed that in some cases the combined-MC solution converges to a non-zero value of P_0 while simulations indicated that the system is unstable. However, all studies that we have performed showed that in the region where the system is stable, the combined-MC approximation is very accurate (see results presented in Sec. 7.6).

Other Approximations

The proposed approach based on the combined Markov chain is scalable. However, the number of slots until reservation success is infinite and the distribution of X_R should be truncated which results in an approximation. Furthermore, even when the distribution of X_R is truncated, the computations are intensive. We therefore propose in this section three different approaches to reduce the computational complexity required to calculate the moments of X_R.

Reduced Markov Chain The number of states in the Markov chain can be reduced by assuming that idle nodes will independently have a non-empty queue at the beginning of a timeslot, and therefore participate in the competition, with probability $1 - P_0$. This assumption therefore ignores the dependency between timeslots of the status of an idle node and we don't need to track in the MC state variable the number of idle nodes with a non-empty queue. That is, the combined MC chain presented in Section 7.4.2 can be replaced with the single dimension Markov chain proposed in Pawelczak *et al.* (2009) which uses the single state variable k, the number of busy nodes. In each timeslot, $N - k$ nodes then participate in the competition with an Aloha access probability given by $p_m = p(1-P_0)$. The distribution of X_R can then be computed as explained in Section 7.4.2 where $P_s(g)$ given in (7.1) is replaced by $(N-k-1)p_m(1-p_m)^{N-k-1}\psi$, the summations in (7.9) and (7.10) are over a single variable, and the modified transition probabilities in (7.9) and steady state probabilities in (7.10) are obtained from the MC in Pawelczak *et al.* (2009). The solution procedure is then the same as the one described in Section 7.4.2. This approximate model, denoted by 'Pawelczak-MC' in the numerical results, reduces the computational complexity since the state space is smaller.

Approximate X_R Distribution We now propose an approximate approach which significantly reduces the complexity associated with finding the distribution of X_R so that it can be used in scenarios with limited computation capacity. The complexity in (7.9) arises from the fact that the reservation success probability is time-varying and has memory. As an approximation, we propose to simply assume that the number of participating nodes during a reservation period is always n. With this assumption, X_{nR}, the reservation time given that the number of participating nodes is n at each timeslot, has a geometric distribution with probability $(1 - H_{s_{max}})P_s^m(k, n|k < s_{max}) + H_{s_{max}} P_s^m(k = s_{max}, n)$, where $H_{s_{max}}$ is the probability of k being equal to s_{max} and is defined as follows :

$$H_{s_{max}} = \sum_{(k,g)|k=s_{max}} \pi_{(k,g)}^N. \tag{7.11}$$

From the combined MC chain presented in Sec. 7.4.2, we can compute $Pr(g = n|g \neq 0)$ $(n = 1, \ldots, N)$, the distribution of the number of nodes participating in the competition excluding the cases where there is no competitors. We can then approximate the distribution of X_R as follows :

$$X_R \approx \sum_n Pr(g = n|g \neq 0) X_{nR}. \tag{7.12}$$

Given that X_{nR} is geometric, it is then easy to compute the moments of X_R. Since P_0 is not known, the same solution procedure as the one described in Section 7.4.2 should be used. This approximate model is denoted by Combined-MC-Dist in the numerical results.

Average X_R Distribution This third approach further simplifies the computations by calculating the reservation time with the average number of idle nodes in competition with a non-empty queue. Let define

$$\bar{G} = \sum_{(k,g)|g\neq 0} g\pi_{k,g}^N. \tag{7.13}$$

where the steady-state probabilities $\pi_{k,g}^N$ are obtained from the combined MC chain presented in Section 7.4.2. Then X_R is approximated with a geometric distribution with probability $(1 - H_{s_{max}})P_s^m(k, \bar{G}|k < s_{max}) + H_{s_{max}} P_s^m(s_{max}, \bar{G})$. Since P_0 is not known, the same solution procedure as the one described in Section 7.4.2 should be used. This approximate model is denoted by Combined-MC-Avg in the numerical results.

7.5 Switching MAC Protocol Delay Analysis

In this section, we provide the delay analysis for the multichannel OSA switching MAC protocol. The same approach as for the delay analysis for the buffering MAC protocol is followed : we first provide a system occupancy Markov chain in Section 7.5.1 and then we derive an Geom/G/1 queue analysis in Section 7.5.2. The only difference between the buffering and switching MAC protocols is that in the switching protocol, if a SU senses that its reserved channel is occupied by a PU, it immediately returns to the group of idle nodes with non-empty queue and participates in the competition, while in the buffering MAC protocol the node would remain on its reserved channel. To facilitate the analysis of the switching MAC protocol, the system occupancy MC and combined MC for service time analysis are embedded after the sensing period. Finally, for sake of brevity, we omit details in the analysis that are similar to the buffering MAC protocol analysis.

7.5.1 Occupancy Markov Chain

The states of the queue occupancy Markov chain are similar to the buffering case and are given by $(\vec{n}, \vec{b})$. However, the observation time where the Markov chain has been imbedded is immediately after the initial sensing period when the status of the channels has already been determined. We therefore have that when $b_i = 1$ the reserved channel for node i is available in this timeslot for transmission (the transmission may however be unsuccessful due to the packet capture model).

Similar to the buffering case, for a state A, we can define k_A as the total number of busy nodes who have a channel and whose channels are available in this timeslot and g_A as the total number of idle nodes with non-empty queue. The switching model facts are very similar to the buffering model with the difference that a node who transmitted on its assigned channel in this timeslot may, if it does not finish its packet transmission, come back to the competition with probability p_c if a PU is detected on its assigned channel. For $N = 1$, the Markov chain is illustrated in Fig. 7.3 where the transitions after states $(2, 0)$ and $(2, 1)$ are not shown. It can be seen that compared to Fig. 7.2 for the buffering policy, there are only two changes : a new transition may occur from state (1,1) to (1,0) if no arrival occurs, transmission is not finished, but the channel is missed in the next timeslot, and to state (2,0) with the same conditions and an arrival occurs.

Transition probabilities for this Markov chain can be found using Alg. 2 with the same set of impossible states (IS). Similar to the buffering model, when a node i is busy ($b_i = 1$), its transitions are independent of other nodes. For $b_i = 0$ and $n_i > 0$, the node participates in competition. Compared to the buffering model, it can be seen that the main change is for the case where there is a transition between $b_i = 1$ and $c_i = 0$ because this transition may occur in the switching model due to both a service termination or a PU positive sensing. Moreover, we can see that the probability of success is updated by $1 - p_c$ because a success in competition should also be followed by the availability of the reserved channel to be a

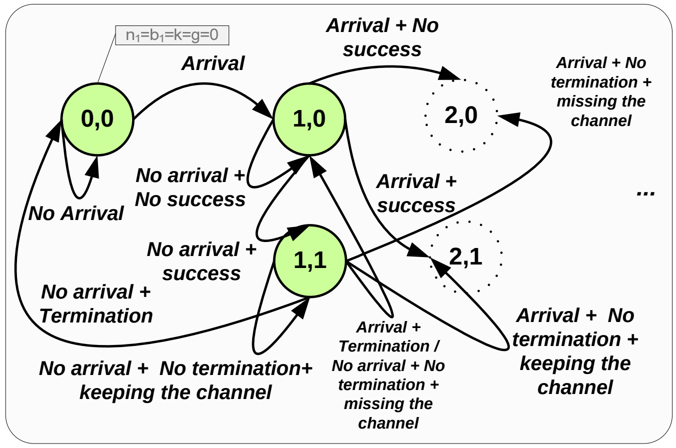

Figure 7.3 An example of the proposed queue occupancy Markov chain for switching policy when $N = 1$ and $M_C > 0$.

reservation success.

For the case where we have $k = s_{max}$, the success depends on the nodes who had a transition from $b_i = 1$ to $c_i = 0$. Any of them may have a transition due to missing the channel and PU arrival (kept in vector P_m) or a termination (kept in vector P_t). P_{mt} keeps the total probability of both events. Then, P_{Allm} is defined to be the probability that all transitions from $b_i = 1$ to $c_i = 0$ have been due to PU arrival. In this special case, no reservation success is possible. If at least one node among them has finished its transmission, with the probability P_{10T}, then a competition success may result into a reservation success with the probability $1 - p_c$.

7.5.2 Service Cycle Analysis

The nodes queue in the multichannel OSA network with the switching MAC protocol can also be modeled with a Geom/G/1 queue. As illustrated in Fig. 7.4, the Geom/G/1 service time in this queue is composed of alternating renewal processes : uncompleted packet transmissions (a part of the packet however can be transmitted) and competitions for channel reservation alternate until the packet is entirely transmitted. Since the packet length is geometric and the occurrence of an interruption is a Bernoulli event, and both are independent, the uncompleted packet transmission time is thus memoryless and identical. The unsuccessful transmission attempts therefore form a renewal process. The reservation periods X_R during the transmission of a packet are also identical and form the second renewal process. Let X be the service time of the Geom/G/1 queueing model, which can be given by

$$X = X_R + L_e + nX_R = L_e + mX_R, \tag{7.14}$$

Algorithm 2 The algorithm to find the transition probabilities (FindProb(A,B)) in the switching model. Note that $f^0_{ap}(n_i, m_i) = \lambda^{m_i-n_i}(1-\lambda)^{1-(m_i-n_i)}$ $(m_i \geq n_i)$ and $f^1_{ap}(n_i, m_i) = \lambda^{1-(n_i-m_i)}(1-\lambda)^{n_i-m_i}$ $(m_i \leq n_i)$.

```
if (k_A == s_max & Num_(t,AB) == 0) then
    AllBusyNoTer = 1
end if
for i = 1 : N do
    if (n_i == 0) then
        P_i = f^0_ap(n_i, m_i)
    else
        if b_i == 1 then
            if (c_i == 1) then
                P_i = (1 − qη)(1 − p_c) f^0_ap(n_i, m_i)
            else
                if (k_A! = s_max) then
                    P_i = (1 − qη)(p_c) f^0_ap(n_i, m_i) + (qη) f^1_ap(n_i, m_i)
                else
                    Put i in S_10
                    h = h + 1
                    P_i = 1
                    P_t,h = (qη) f^1_ap(n_i, m_i)
                    P_m,h = (1 − qη)(p_c) f^0_ap(n_i, m_i)
                    P_mt,h = (1 − qη)(p_c) f^0_ap(n_i, m_i) + (qη) f^1_ap(n_i, m_i)
                end if
            end if
        else
            Put i in S_g
            P_i = f^0_ap(n_i, m_i)
        end if
    end if
end for
for i = 1 : |S_10| do
    P_10T = P_10T + ∏_1^{h−1} P_m,h P_t,h ∏_{h+1}^{|S_10|} P_mt,h
end for
P_Allm = ∏_1^{|S_10|} P_m
FindProb(A,B)=∏_i^N P_i ∏ P_mt,h
***** Handling Competitions ***** (IF S_g ≠ ∅)
if (k_A! = s_max) then
    Find j s.t. (c_{S_g[j]} == 1)
    if j > 0 then
        P_{S_g[j]} = P_{S_g[j]} ∗ (P_s(|S_g|) / |S_g|)(1 − p_c)
    else
        if (AllBusyNoTer == 0) then
            P_{S_g[1]} = P_{S_g[1]} ∗ [1 − P_s(|S_g|)(1 − p_c)]
        end if
    end if
else
    Find j s.t. (c_{S_g[j]} == 1)
    if j > 0 then
        P_{S_g[j]} = P_{S_g[j]} ∗ (P_s(|S_g|) / |S_g|)(1 − p_c)
        P_{S_m[1]} = P_10T
    else
        if (AllBusyNoTer == 0) then
            P_{S_g[1]} = P_{S_g[1]} ∗ [P_Allm + P_10T(1 − P_s(|S_g|)) + P_10T P_s(|S_g|)(1 − p_c)]
        end if
    end if
end if
FindProb(A,B)=∏_i^N P_i
```

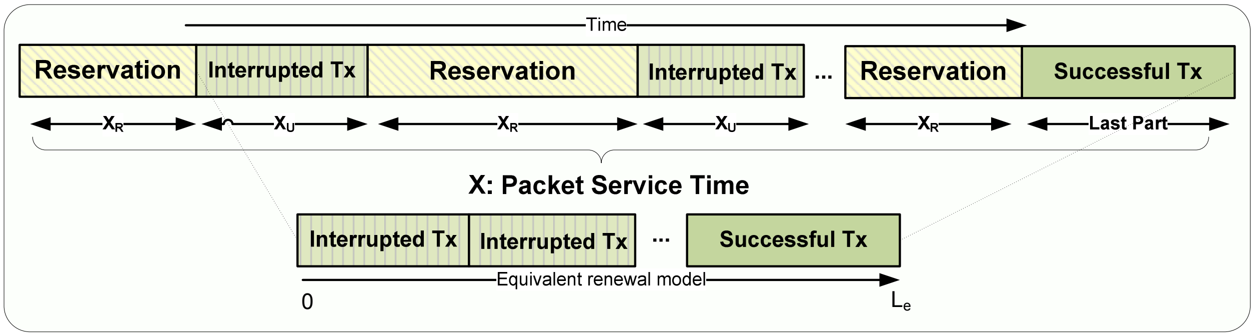

Figure 7.4 Unsuccessful transmission and reservation periods create alternating renewal processes.

where L_e is the enlarged packet length (in number of slots), which is geometrically distributed with the probability $q\eta$, and $n = m - 1$ is the number of switching during the transmission of the packet. To find the service time, we must find the distribution of X_R and the number of renewals n, which are addressed in Section 7.5.2 and Section 7.5.2, respectively.

Reservation Period

The approach to find the distribution of X_R for the switching MAC protocol is similar as the one followed for the buffering model in Sec. 7.4.2. The state variable of the combined Markov chain is still given by a 2-tuple (k, g), however, the observation time where the combined Markov chain is embedded for the switching MAC protocol is after the initial sensing period at the beginning of each timeslot such that the status of the channels for the current timeslot is known. k is therefore the number of reserved channels on which there will be a transmission attempt in this timeslot and g is the number of idle nodes with a nonempty queue which participate in the competition on the control channel for reservation. The number of idle nodes with an empty queue is thus given by $N - g - k$.

To find transition probabilities, we use four auxiliary variables i, j, n and m. Note that in the following, k and g refer to $k(t)$ and $g(t)$ for the current timeslot t. Among k busy nodes, i nodes independently finish the transmission of their packets, each with the probability $q\eta$. We use the notation $T_k^i = \binom{k}{i} q\eta^i (1 - q\eta)^{k-i}$ for the probability of this event. Any of those i nodes may join in the next timeslot, with probability $1 - P_0$, the idle nodes with a nonempty queue. Otherwise, they join the idle nodes with an empty queue. The probability that n nodes out of the i finishing nodes will be among the $g(t+1)$ nodes in the next timeslot is thus given by $\binom{i}{n} {P_0}^{i-n} (1 - P_0)^n$. For the remaining $k - i$ nodes which do not finish their transmission, j nodes may leave their channels due to PU presence sensing with probability $F_{k-i}^{k-i-j} = \binom{k-i}{j} 1 - {p_c}^{k-i-j} (p_c)^j$. Those nodes also join the $g(t+1)$ idle nodes with a nonempty queue after the sensing period. The remaining $k-i-j$ nodes stay on their channels, so they are among the $k(t+1)$ nodes.

Considering g participants in competition, we have the same results for competition success as for the buffering case. One node may then be successful and join the $k(t+1)$ nodes with probability $P_s(k,g)(1-p_c)$. The $1-p_c$ term is included because the node may be successful in competition and reservation during this timeslot, but the reserved channel might be unavailable in the next timeslot. Therefore, without starting the transmission, the node has to come back to competition. All other $g-1$ nodes remain in the group of idle nodes with a nonempty queue, i.e., they are among $g(t+1)$ nodes. Finally, any of $N-k-g$ idle nodes with an empty buffer may receive a packet and join the participants in competition in the next timelot. The probability to have m such nodes is given by $\binom{N-k-g}{m}\lambda^m(1-\lambda)^{N-k-g-m}$.

We can thus, for the switching MAC protocol, express the transition probabilities of the combined Markov chain in the form $P_{(k,g)->(z,h)}$ as follows :

where $Y_a^{b,c}$ is given in (7.8). Given the transition probabilities for the combined Markov chain for the switching MAC protocol, the distribution of X_R can be obtained as explained in Section 7.4.2. The second and third approximations given in Section 7.4.2 to find the distribution of X_R can also be used (the approximation based on the model presented in Pawelczak *et al.* (2009) is not applicable for the switching MAC protocol).

Renewal Process

As illustrated in Fig. 7.4, X_U, any part of the whole service time where the node has a reserved channel, is the inter-arrival time of the channel unavailability events. Let us use the random variable A_u to denote the inter-arrival time of those events. A_u has a geometric distribution with probability p_c. From the Renewal Theory point of view Cox (1962), $n = m-1$ is the number of renewals from 0 to L_e where the process which is renewed is A_u. The number of renewals can be found from Renewal Theory results Cox (1962).

However, as we have geometric distributions for both L_e and A_u, we can use the following simpler approach. Due to the memoryless packet length, the remaining part of the packet in any operating period is geometrically distributed with parameter q. The total number of renewals is thus the number of trials until success, which is a geometric distribution of type-II, with the probability of success in a period being $Pr(L_e <= A_u)$, the probability of sending the whole packet in the given period. However, the number of reservation periods which appear in the calculation of service time is $m = n+1$ because the first reservation period is not counted as a renewal (see Fig. 7.4). We thus have :

$$Pr(n == k) = (1 - Pr(L_e <= A_u))^k Pr(L_e <= A_u). \tag{7.15}$$

$$P_{(k,g)->(z,h)} = \begin{cases} 0 & \text{if } z > k+1, \forall g, h, \\ 0 & \text{if } \forall k, z, h < g-1, \\ T_k^{(0)} F_k^k S_g^{(1)} (1-p_c) Y_{(h-g+1)}^{(k-z+1),(N-k-g)} & \text{if } z = k+1, \\ \sum_{i=0}^{k-z} T_k^i F_{k-i}^z (1 - S_g^{(1)}(1-p_c)) Y_{(h-g-(k-i-z))}^{i,(N-k-g)} + \sum_{i=0}^{k-z+1} T_k^i F_{k-i}^{z-1} (S_g^{(1)}(1-p_c)) Y_{(h-g-(k-i-z))}^{i,(N-k-g)} & \text{if } 0 < z \leq k, k+z \neq 2s_{max}, \\ \sum_{i=0}^{k-z} T_k^i F_{k-i}^z Y_{(h-g-(k-i-z))}^{i,(N-k-g)} + \sum_{i=0}^{k-z+1} T_k^i F_{k-i}^{z-1} (S_g^{(1)}(1-p_c)) Y_{(h-g-(k-i-z))}^{i,(N-k-g)} & \text{if } z = k = s_{max}, \\ \sum_{i=0}^{k-z} T_k^i F_{k-i}^z (1 - S_g^{(1)}(1-p_c)) Y_{(h-g-(k-i-z))}^{i,(N-k-g)} & \text{if } z = 0, \end{cases} \tag{7.15}$$

For any two type-I geometrically distributed random variables R_1 and R_2 with parameters p_1 and p_2, we can show that

$$Pr(R_1 <= R_2) = \frac{p_1}{1 - (1-p_1)(1-p_2)}. \tag{7.16}$$

We thus have :

$$Pr(L_e <= A_u) = \frac{q\eta}{1 - (1-q\eta)(1-p_c)}, \tag{7.17}$$

$$E[n] = \frac{1 - \frac{q\eta}{1-(1-q\eta)(1-p_c)}}{\frac{q\eta}{1-(1-q\eta)(1-p_c)}} = \frac{p_c(1-q\eta)}{q\eta}, \tag{7.18}$$

and

$$E[n^2] = (E[n])^2 + \frac{1 - \frac{q\eta}{1-(1-q\eta)(1-p_c)}}{[\frac{q\eta}{1-(1-q\eta)(1-p_c)}]^2}. \tag{7.19}$$

To find the second moment of X it should be taken into account that we have a random sum and the two parameters n and L_e are not independent. Using the following facts :

$$\sum_{k=1}^{\infty} k(1-q)^{k-1} = \frac{1}{q^2}, \tag{7.20}$$

$$\sum_{k=1}^{\infty} k(k-1)(1-q)^{k-2} = \frac{2}{q^3}, \tag{7.21}$$

and thus

$$\sum_{k=1}^{\infty} k^2(1-q)^{k-2} = \frac{2-q}{q^3(1-q)}, \tag{7.22}$$

the term $E[L_e n]$ can be given by :

$$E[L_e n] = \\ \sum_{k=1}^{\infty} E[L_e n|(L_e = k)]Pr(L_e == k) = \sum_{k=1}^{\infty} kE[n|(L_e = k)]Pr(L_e == k) \\ = \sum_{k=1}^{\infty} k(k-1)p_c q\eta(1-q\eta)^{k-1} = \frac{2p_c(1-q\eta)}{(q\eta)^2}. \tag{7.23}$$

We have used the fact that the number of renewals in the interval $[0\ k]$ of a type-I geometric with the parameter p_c is equal to $(k-1)p_c$.

From the moments of n and X_R, it is then straightforward to find the moments of X and solve the Geom/G/1 queue. Note that since P_0 is initially unknown and is required to find the distribution of X_R, the solution procedure outlined in Section 7.4.2 can be followed. The initial value of P_0 is computed with the lower bound of X_R obtained for the case where only a single node is competing and for which $E[X_R] = (p\chi(1-p_c))^{-1}$.

7.6 Simulation and Numerical Results

In this section, we verify the accuracy of the proposed models and approximations by comparing simulation results with numerical evaluations of the theoretical results presented in this paper. We also study the impact of several parameters on the multi-channel OSA MAC protocols performance. Simulations are done with Matlab where each simulation scenario is run for 350,000 timeslots and the same scenario is repeated 10 times. Unless indicated otherwise, we have $M_C = N$ data channels. In Sec. 7.6.1 to 7.6.4, we investigate the buffering MAC protocol while in Sec. 7.6.5 we present results for the switching MAC protocol.

7.6.1 Impact of Arrival Rate

Fig. 7.5 shows the average system time as a function of the arrival rate for the buffering MAC protocol and for different system parameter values. It can be first observed that, as expected, the queue occupancy Markov chain provides accurate results compared with simulation results, called 'truncated simulation', obtained with the same limited buffer size assumed to numerically solve the occupancy Markov chain. All approximations also provide acceptable results at low arrival rate while for larger values, when the system gets close to instability, their accuracy decreases. For example, for $N = 10$ users, the combined-MC approximation is accurate within 3% for a normal operation region below 80% of maximum loading. It was observed through several scenarios that, as expected, the combined Markov chain provides the best approximation while the average based and distribution based approximations both diverge sooner than the two other schemes when the queue approaches the instability region. In the following, to help the figures readability, we will only present some of the results obtained with the different approximations.

7.6.2 Impact of Medium Access Probability

The probability of medium access on the control channel has a major influence on the performance of the network. A very small p or a very large p makes the system unstable because, in the former case, a packet has to wait several slots before the node tries to reserve a channel, and, in the latter case, there will be a lot of collisions such that the reservation success probability will be very low. In Fig. 7.6, we can see that for any given scenario of the buffering MAC protocol, there is an optimal value for p which depends on the arrival rate. This optimal value decreases when the total traffic load on the control channel increases due to either an increase in the arrival rate, an increase in the number of nodes or smaller packets. Interestingly, it can be observed that by increasing the packet length (smaller q), even though the total system load and the delay increase because the packets are now longer, the optimal probability of channel access p also increases. This is due to the fact that with longer packets, the nodes return less frequently to the control channel. There is thus less competition for channel reservation and a higher probability of control channel access is optimal.

It is also observed that the optimal control channel access probability is significantly larger than for the saturated traffic model where the optimal value of p is very small Pawelczak *et al.* (2009). There is also a large range of values of p where the performance is similar. However, this range decreases with an increase in traffic load of the system.

7.6.3 Impact of Packet Length

As observed in Fig. 7.6, the system time increases when the system load increases either due to an increase of arrival rate or packet length. In Fig. 7.7, we investigate the impact of both parameters on the buffering MAC protocol when the traffic load λ/q is fixed. It can be observed that as the packet length increases (and the arrival rate decreases), the system time decreases. This is due to the fact that in the network model with an Aloha-type medium access algorithm, the main bottleneck is the competition on the control channel to reserve a data channel. For a fixed traffic load, when the packet length increases, fewer nodes will therefore participate in the competition and the nodes with a reserved channel spend more time on the reserved channel and less time waiting on the control channel to get a reservation success.

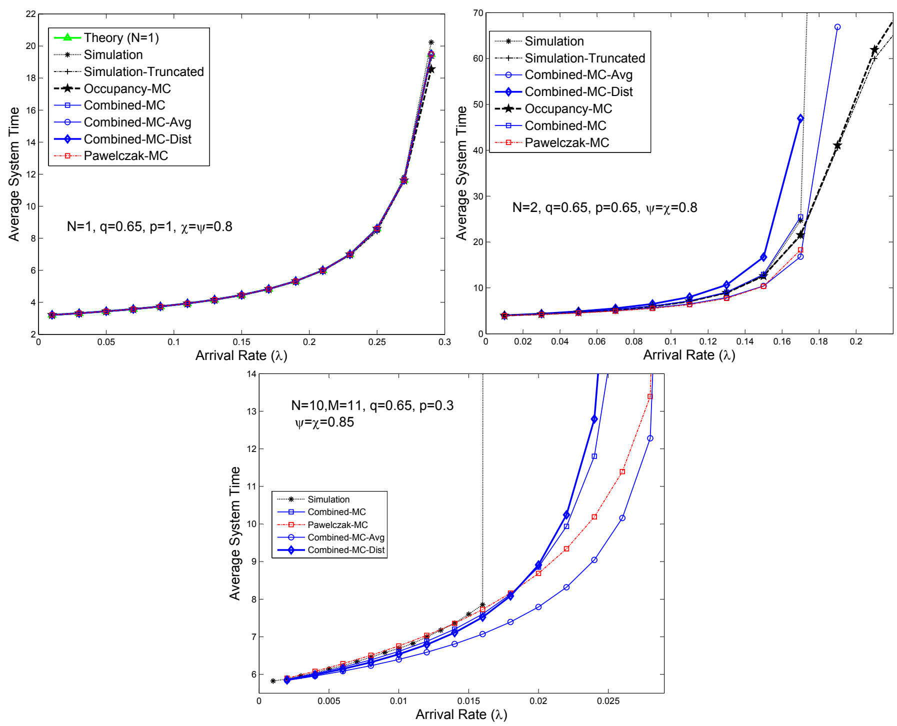

Figure 7.5 Performance comparison of the proposed schemes versus the variation of the arrival rate for the buffering MAC protocol.

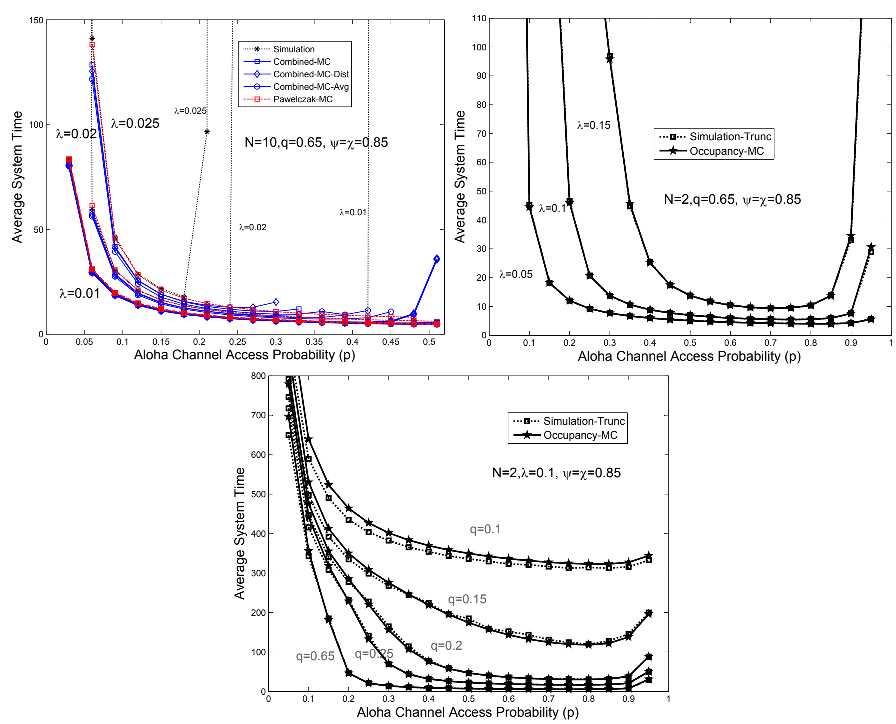

Figure 7.6 Average system time for the buffering MAC protocol versus the Aloha-type probability of control channel access, for different arrival rate and packet length values.

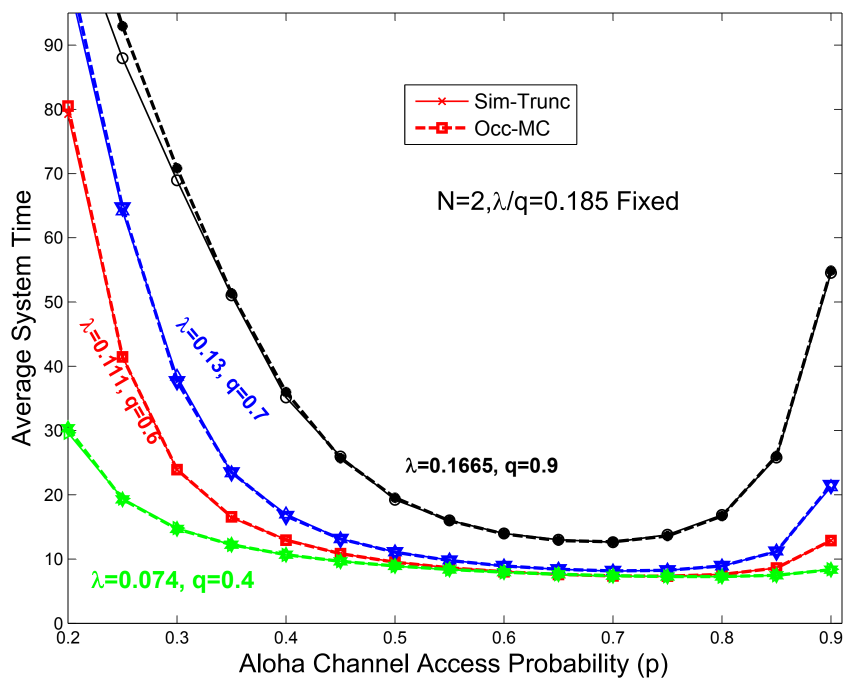

Figure 7.7 Performance comparison for the buffering MAC protocol for a fixed traffic load of $\lambda/q = 0.85$ and $\chi = \psi = 0.85$.

7.6.4 Impact of Transmission Capacity

We investigate in this section the impact of transmission capacity which depends on the number of available channels and the channel availability ψ. In Fig. 7.8, the number of nodes is fixed to $N = 10$ and the number of data channels is increasing from $M_C = 1$ to $M_C = 10$, for two different values of packet length. As discussed previously, when the number of nodes is high the probability of success in competition decreases and all nodes spend a longer time for reserving a channel. Even though there are available channels, no node can make a reservation ; therefore, the increase of the number of channels provides no gain for the network. When the packet length is small ($q = 0.65$), the packet transmission time is short, so a winner again joins quickly the competing users and the probability to have even two channels busy at the same time is low. With a longer packet length ($q = 0.065$), fewer users compete and data channels are more utilized. The system time therefore decreases until 6 data channels are available and marginal gains are achieved for larger values of M_C.

The observation that the main bottleneck of an Aloha-based cognitive radio MAC protocol is the channel reservation competition on the control channel leads us to propose the concept of channel and node clustering. That is, instead of having a single control channel the CR nodes and channels can be divided into clusters where each cluster has its own control channel.

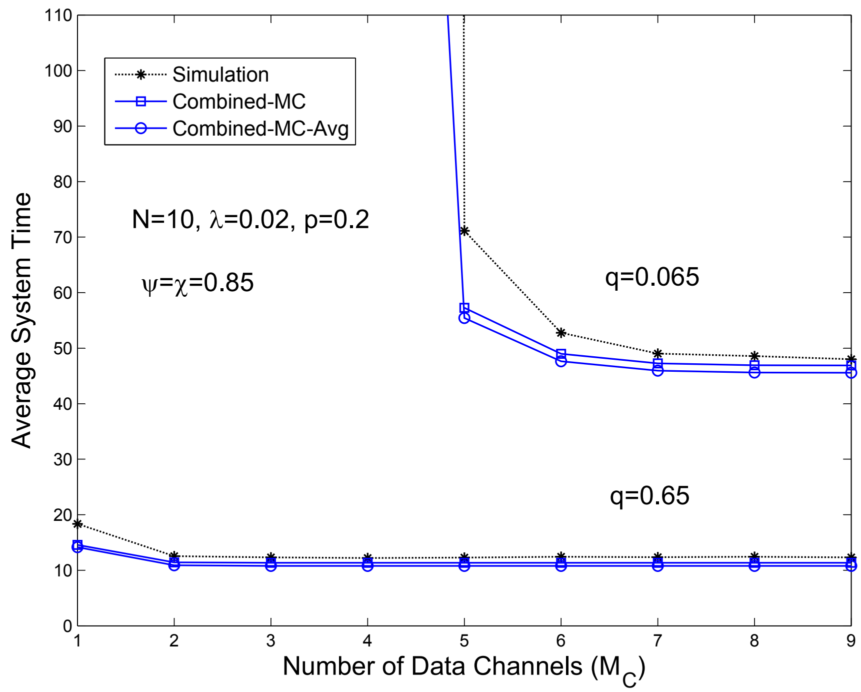

Figure 7.8 System time as a function of the number of data channels M_C for the buffering MAC protocol.

The results presented in Fig. 7.9 show the performance of such a clustered system. We assumed a network with 10 nodes and 10 channels and compared the performance when there is a single control channel and all nodes operate under a single entity of the buffering MAC protocol (i.e., $N = 10$ and $M_C = 9$) with the case where the nodes and channels are divided into two clusters and each cluster has its own control channel and operate under an independent entity of the buffering MAC protocol (i.e., $N = 5$ and $M_C = 4$). The probability of medium access, p, has been adjusted to the value which minimizes the delay for each case. It can be observed that even though one additional channel is now used as a control channel in the two-cluster case, and we thus have fewer data channels, the system time has significantly decreased for all values of arrival rate. Furthermore, a larger arrival rate can be served when there are two clusters. Those results confirm our hypothesis that the delay bottleneck is due to the reservation competition on the control channel and the fact that, depending on the system parameters, it might be more beneficial to add control channels than to add data transmission capacity.

Another important issue is the trade-off between the number of channels and their availability, when, for instance, a cognitive radio network has the option of utilizing a spectrum with a large number of channels with a lower availability per timeslot and another spectrum with a smaller number of channels with a higher availability. In Fig. 7.10, we investigate this

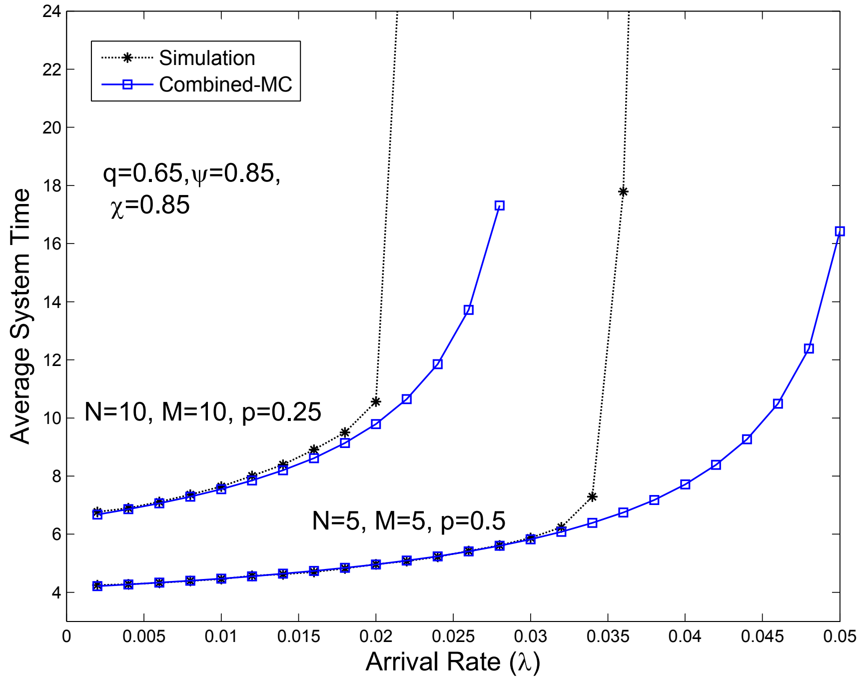

Figure 7.9 Performance comparison for the buffering MAC protocol of a network with $M = N = 10$ and the clustered network with two clusters of $N = M = 5$.

tradeoff where the channel availability ψ decreases when M_C increases. Note that we kept the control channel availability constant. The interesting point to observe is that there is an optimal point where the average delay is minimized. This is due to the fact that because of the control channel access bottleneck, an increase in the number of channels does not provide any significant gain for the network (see results presented in Fig. 7.8) while the decrease in availability of the assigned channel increases the transmission time of the packet. Such an optimal point exists for any network as a function of the system parameters, so the analytical queueing results presented in this paper enable the network designer to select the best operating channel set.

7.6.5 Switching MAC Protocol

In this section we investigate the performance of the switching MAC protocol. Fig. 7.11 shows the system time as a function of the arrival rate for $N = 10$ nodes. It can first be observed that the analytical approximations provide a good prediction of the system performance. Those results also show that there is a significant performance degradation versus the results presented previously for the buffering MAC protocol. For similar parameters, we observe approximately 50% lower delay in the buffering case. This is due to the fact that for a packet transmission a node has to incur several reservation periods, which are the delay

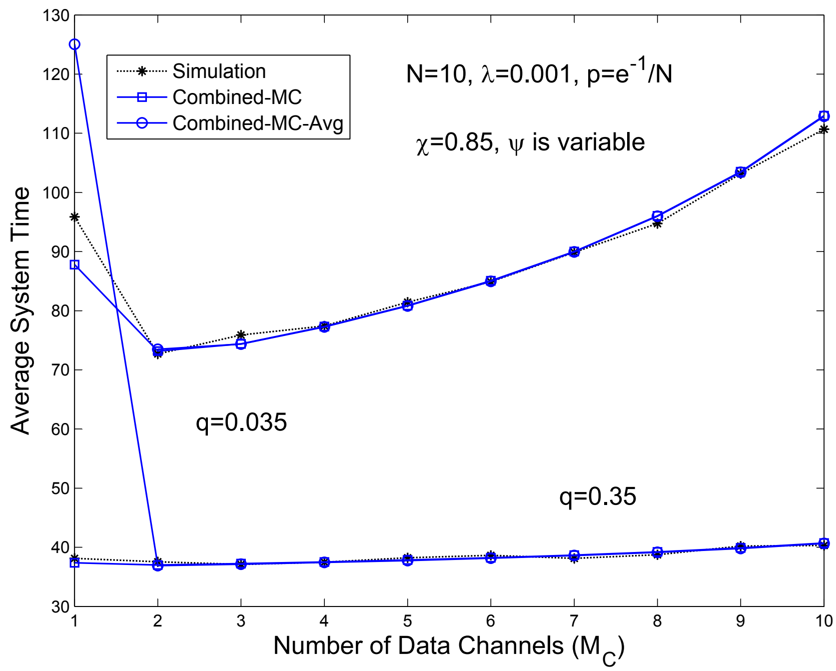

Figure 7.10 Impact of the variation of the number of data channels when the channel availability, ψ, decreases for the buffering MAC protocol. For $M_C = [1 \ldots 10]$, we have $\psi = [0.85, 0.8, 0.75, 0.7, 0.65, 0.6, 0.55, 0.5, 0.45, 0.4]$.

bottleneck, and furthermore this creates more competition on the control channel, which further increases the reservation delay.

In Fig. 7.12, the performance versus the variation of the number of data channels and channel availability ψ is illustrated. It can easily be seen that the performance of the switching policy significantly degrades especially when the packet length is long. This is due to the fact that the longer the packet are and the more frequent the transmission occurs, the more often the nodes return to the control channel and wait to get a reservation. However, we can still observe for large packets ($q = 0.035$) an initial improvement when the number of channels increases from one to two.

Fig. 7.13 shows the impact of Aloha access probability on the packet system time for the switching MAC protocol. Those results first validate the occupancy Markov chain model. It is also interesting to observe that even though the switching policy is more sensitive to an increase in packet length due to the frequent returns to the control channel, there are some scenarios, as the one illustrated in the figure, in which having a lower arrival rate with longer packets provides a better performance than having a higher arrival rate of shorter packets.

Since the switching occurs due to PU arrivals and not fading, we investigate the impact of variation of p_c on the system time. The results for the switching and buffering MAC protocols

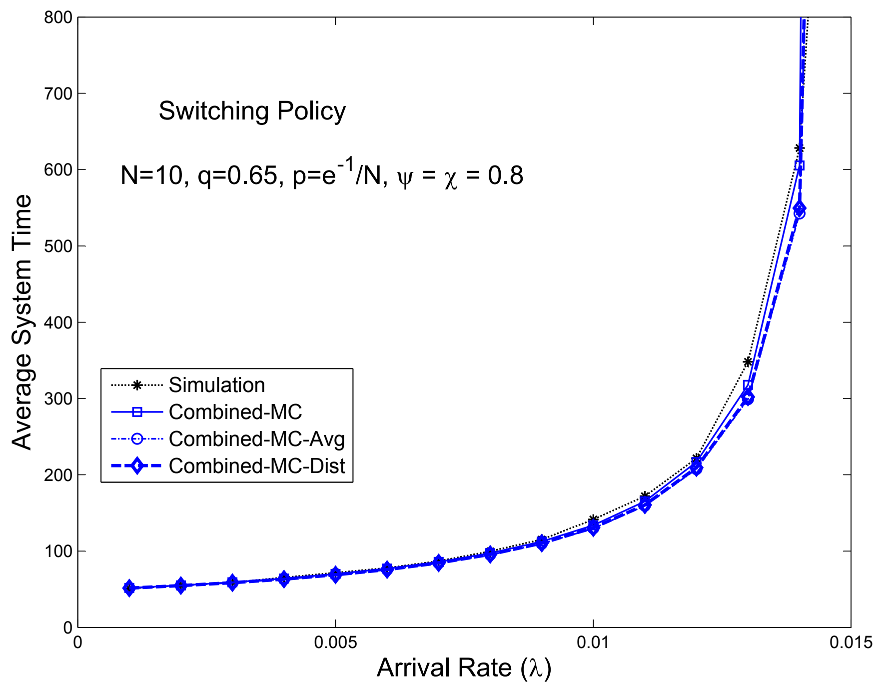

Figure 7.11 System time for the switching MAC protocol as a function of the arrival rate.

are illustrated in Fig. 7.14. When $p_c = 0$, the two policies are naturally the same. When p_c increases, the performance of the buffering MAC policy only slightly deteriorates due to a small increase in packet transmission time. However, for the switching MAC policy, each PU interruption incurs a large penalty of channel reservation. Therefore, the performance quickly degrades as a function of p_c. It is worth mentioning that we have also ignored the switching time required to align the radio in the switching policy Park *et al.* (2011). Naturally, assuming a switching time will further degrade the performance of the switching policy.

Comparing the results of the buffering and switching policies reveals that in all scenarios, the buffering MAC protocol outperforms the switching. This is simply due to the homogeneity of the channels and memoryless PU presence. That is, with the switching policy if the channel is not available, the node tries to reserve a channel. In the best case, the reservation process will last one timeslot. However, the reserved channel has the same probability of being available in the next timeslot as if the node stayed on the same channel with the buffering policy. So there is no way for a node to decrease its transmission delay by switching to another channel.

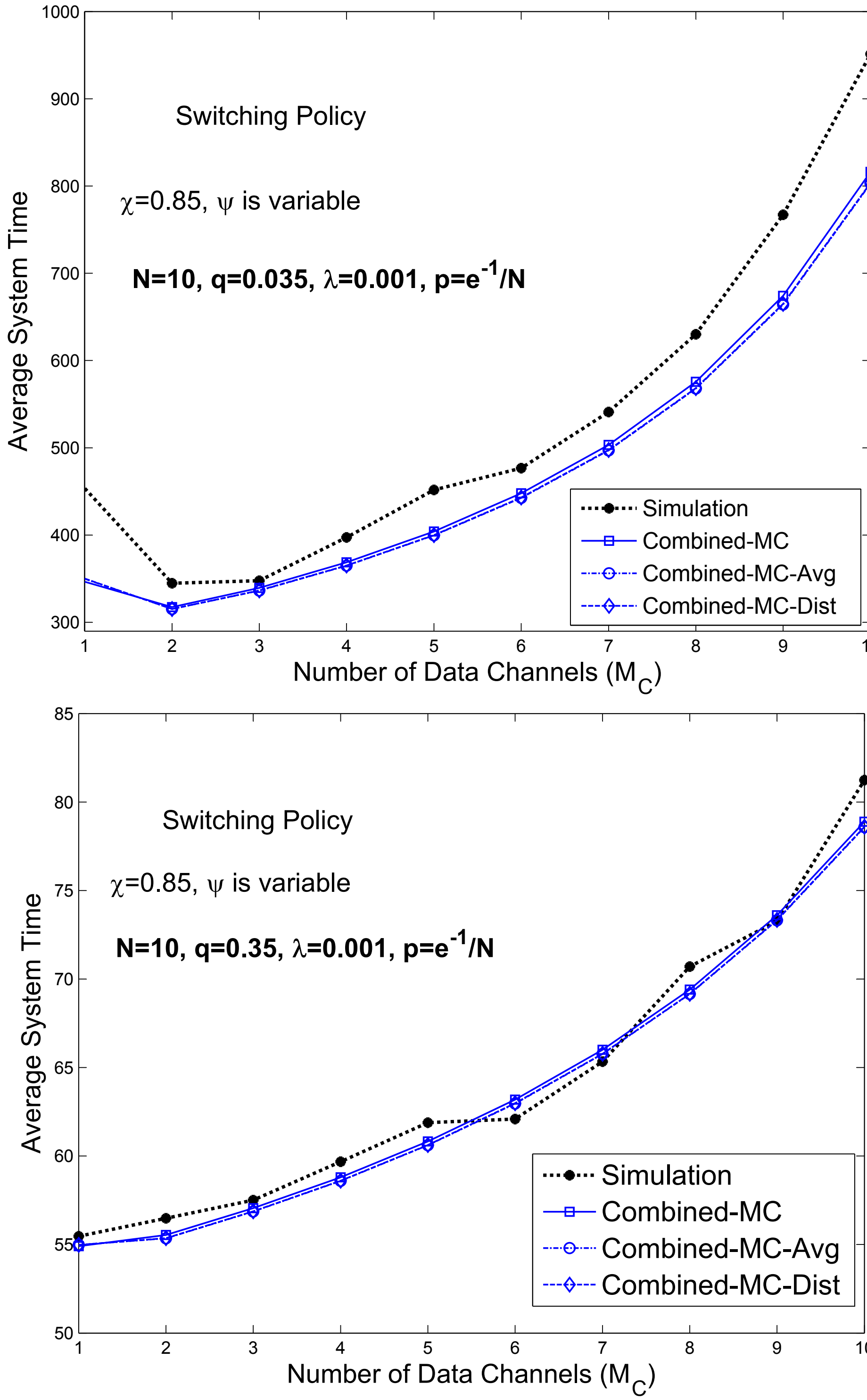

Figure 7.12 Performance comparison versus the variation of the number of data channels M_C when the channel availability, ψ, decreases for the switching MAC protocol. For $M_C = [1 \ldots 10]$, we have $\psi =$[0.85, 0.8, 0.75, 0.7, 0.65, 0.6, 0.55, 0.5, 0.45, 0.4].

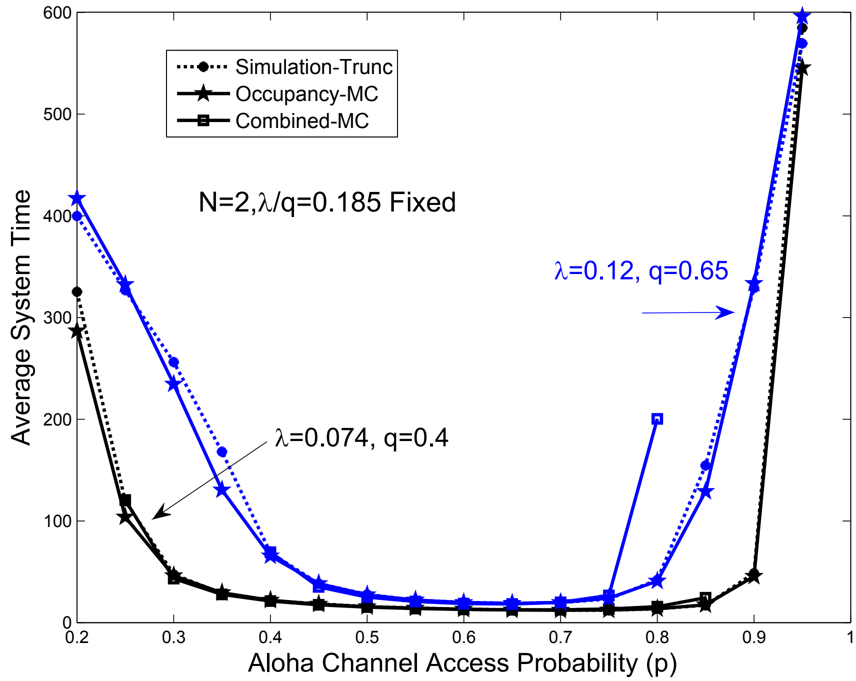

Figure 7.13 Average system time for the switching MAC protocol versus the Aloha-type probability of control channel access. ($M_C = N, \chi = \psi = 0.8$).

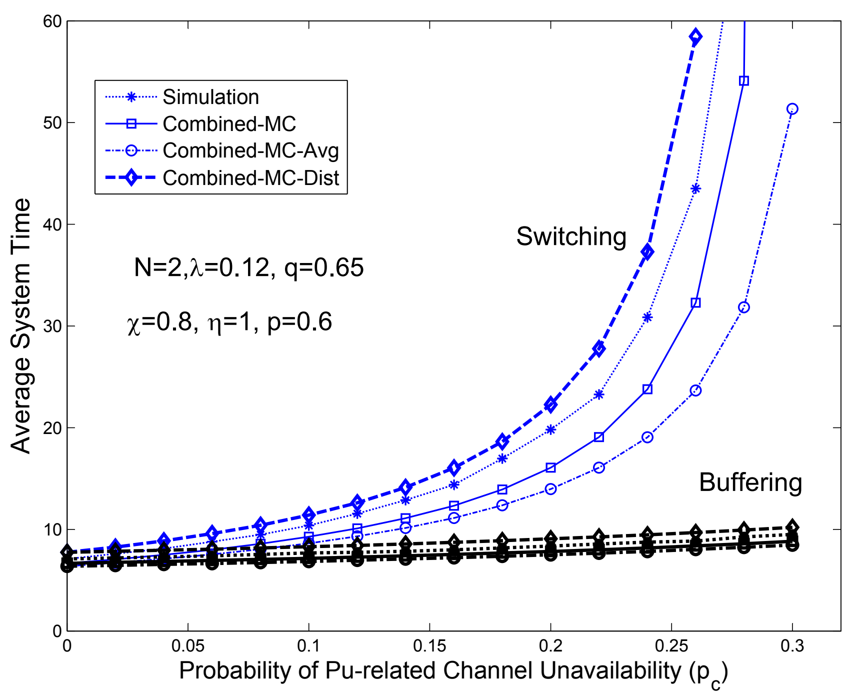

Figure 7.14 Average system time for the switching and buffering MAC protocols as a function of PU probability of activity (p_c).

7.7 Conclusion

In this paper, a delay and queueing analysis for a multi-node network with an Aloha-type medium access model was provided. Two different recovery models were considered : a waiting and buffering recovery policy where the CR node waits for the primary user to vacate the channel and continues the transmission on the same channel, and a switching policy where after the appearance of primary user, a spectrum handover occurs. It was observed that access to the control channel to reserve a data channel is the major bottleneck in Aloha, so any approach which decreases the number of competitors, such as having fewer but longer packets and node clustering, improves the performance. The probability of medium access in Aloha should also be adjusted carefully to have the minimum delay. With the assumption of having homogenous and memoryless channels, a buffering policy always outperforms the switching policy. An important area of future work would be to improve the channel occupancy model with a Markov chain channel instead of the assumption of independent availability per timeslot. Furthermore, since the Aloha access to the control channel represents a major contributor to the delay, alternative control channel strategies should also be investigated following the methodology presented in this paper.

CHAPTER 8

ARTICLE 6 : A HISTORY-AWARE GREEDY CHANNEL RESTORATION SCHEME FOR COGNITIVE RADIO-BASED LTE NETWORKS

Arash Azarfar, Jean-François Frigon and Brunilde Sansò
Published in Elsevier Physical Communication, March 2014
Department of Electrical Engineering
École Polytechnique de Montréal
C.P. 6079, succ. centre-ville, Montréal, QC, Canada, H3C 3A7
Email : {arash.azarfar, j-f.frigon, brunilde.sanso}@polymtl.ca

Cognitive radio-based Long Term Evolution (LTE) networks benefit from the powerful features of cognitive Radios (CR), such as learning and reconfigurability, enabling them to perform, if required, channel switching to a channel with higher quality. This process of searching and sensing other channels is a restoration (recovery) process where the objective is to find the best channel in the shortest time. We propose in this paper a history-aware greedy restoration scheme triggered not only when the quality of the current operating channel of the user goes below a threshold, but at regular intervals. Based on the state of the current channel, our scheme computes the optimal number of channels to be sensed in this restoration period and this number is dynamically updated after each channel sensing result. Intrinsic features of learning and history-awareness of CRs are used to create the list of channels to be sensed based on the channels' background and historical information. The sensing order improves the restoration mechanism by providing a shorter restoration time or a restored channel with a higher quality. We show that the proposed combined scheme provides improvements for the CR-based LTE network's throughput, compared to other restoration schemes which work based on only greediness or history-awareness.

keywords Opportunistic Spectrum Access (OSA), Cognitive Radio (CR), Long Term Evolution (LTE), Learning, History-awareness, Channel restoration.

8.1 Introduction

Interesting and powerful features of cognitive radios (CR) have made them a technology of choice for next generation wireless networks Akyildiz *et al.* (2006). Long term evolution

AZARFAR, A., FRIGON, J.-F. and SANSO, B. (2013) A history-aware greedy channel restoration scheme for cognitive radio-based LTE networks. Elsevier Journal of Physical Communication, special issue on Cognitive Radio in LTE networks, 10, 116-126.

(LTE) as the core of the next generation of cellular networks is not deprived of this promising technology. Consequently, CR-based LTE networks (CR-LTENs) are LTE networks where user equipments (UEs) and base stations (eNBs) are equipped with cognitive features and reconfigurable architectures. In the rest of the paper, the term CR user (node) and LTE user equipment (UE) will be used interchangeably.

During a call, an LTE user equipment may experience transmission quality fluctuations in its operating channel due to several causes such as fading and interference. Specially, when channel aggregation of different channel types (license free, licensed to the LTE network or licensed to a different network) and in different bands is employed, the probability of such fluctuations is higher as the licensees (a.k.a, primary users) may return to some parts of the channel. When, for any reason, the quality of the channel goes below a threshold or the UE must vacate the channel to respect the priority of the licensees, channel switching is unavoidable. Channel switching (a.k.a, spectrum handover) can be modeled as a failure Azarfar *et al.* (2012c) and, when it occurs, the UE in conjunction with the eNB spends an amount of time, called the *restoration time*, to exclude the low quality or occupied sub-carriers from the logical channel and add some new ones to its operating channel, or entirely switch to a new channel. *Recovery or restoration* is thus the process of decision-making to search the spectrum, to evaluate other channel's quality/availability and in case of finding a better channel, to change the channel to a new channel (spectrum handover). The restoration objective is to maximize the overall throughput of the UE, which is equivalent to finding the best channel(s) in the shortest time.

In this paper, building on our previous research work Azarfar *et al.* (2011b,a) where we respectively discussed a history-aware restoration framework and a greedy and intelligent restoration scheme, we propose a history-aware and greedy channel restoration scheme which can be employed in any dynamic resource management framework for LTE networks to increase the overall throughput. Increasing the throughput can be fulfilled through decreasing the restoration time and/or increasing the quality of the selected channel after the restoration. For such a goal, one possible approach is to use more intelligent algorithms for channel selection. We focus on this approach and discuss how the history-awareness and learning capabilities of CR LTE users (eNBs and UEs) enable them to use history-aware (HA) channel selection schemes. That is, by saving and remembering the results of channel sensing, or the spectrum information that they receive from other nodes or central entities, CR nodes can sort the list of potentially available channels to increase the probability of finding a better channel in a shorter time. On the other side, it is mostly assumed in the literature that a restoration scheme is triggered when the quality (i.e., effective capacity) of the operating channel goes below a threshold. We propose in this paper a greedy and intelligent restoration scheme which is triggered periodically, not only when the quality of the channel is not acceptable, to provide higher overall throughput. Therefore, the main contribution of this paper is to propose a history-aware greedy restoration scheme which merges those two concepts together and to discuss the best way to use them in LTE networks in order to improve the overall throughput of a CR-based LTE network.

In the literature, several papers discussed the role of learning and history-awareness in CR-based networks Clancy *et al.* (2007); Zhao *et al.* (2007); Kim et Shin (2008a); Vučević *et al.* (2011); Berthold *et al.* (2008). In Clancy *et al.* (2007), a general model to utilize learning in cognitive radio networks is discussed. In Zhao *et al.* (2007); Kim et Shin (2008a), the long-term statistics of channels' availability create a knowledge base for the CR users to estimate the availability of a channel. The authors of Berthold *et al.* (2008) propose a learning-based scheme for detection of spectrum opportunities in OFDM-based cognitive radio networks. The main difference of our work compared with those previous papers is that our model is more general and applicable to any MAC protocol. Moreover, our learning framework is applicable to multi-user networks with multi-state channels, as in LTE networks, while previous papers in the literature have mostly focused on single-user networks or two-state channel models. In a recent paper Vučević *et al.* (2011), the authors study the coexistence of LTE and Universal Mobile Telecommunications System (UMTS) technologies where learning is employed in a joint resource management framework. Their objective is also to increase the throughput of a mobile node by periodically performing intelligent restoration. However, learning is only employed in this case to make the optimal decision between the access to one of two different access technologies, while our approach is more general.

The rest of the paper is organized as follows. In Section 8.2, we briefly present our channel model and review a general restoration scheme in CR-LTENs. In Section 8.3, we first review the general proposed framework for history-aware (HA) channel selection schemes and then discuss a greedy restoration scheme for an LTE network with homogeneous and heterogeneous channels. Then, their combination is discussed. We explain in Section 8.4 how the proposed scheme can be applicable to LTE-Advanced networks. Simulation results to investigate the performance of the proposed scheme are provided in Section 8.5. Finally, Section 8.6 concludes the paper with some remarks on future research directions.

8.2 System Model

8.2.1 Channel Model

As illustrated in Figure 8.1, considering different levels of interference or fading, a channel from the service capacity point of view (i.e., throughput or rate) can be modeled with a

Markov chain with S states (Z_j, $j = 0, 1, ..., S-1$). The service capacity of the channel in each state j, B_j, is a function of the maximum nominal service capacity (e.g., bandwidth) and the current state Zhang et Kassam (1999). In a deep fade or under intolerable interference or when the channel is a licensed channel and is re-occupied by the primary users, it is assumed that the channel is in state zero ($Z = 0$) and its service capacity is also zero ($B_0 = 0$), which means the channel can not be employed by the CR-LTEN. A channel is more desirable when it is in a state with a larger index, that is, $B_0 < B_1 < ... < B_{S-1}$.

From the aggregation point of view, the available bandwidth for a UE can be modeled with another Markov chain with G states where in state zero, the available bandwidth is 0 and in state G-1, the available bandwidth is 100MHz by aggregation of five 20 MHz-channels. Consequently, the logical channel of an UE can be modeled with a Markov chain with GS states. State $GS-1$ represents the case where all five aggregated operating channels are in state S-1 (maximum capacity).

Nevertheless, in this paper, we focus on a single-channel communication model (no aggregation) where the UE switches to a new channel when the channel cannot be used any more or a channel with a higher throughput is required. As explained in Doyle *et al.* (2011), such a single channel communication model is applicable to the case where channel aggregation is done at the MAC layer and each carrier is thus treated as an individual PHY. For more general cases, our proposed scheme can be embedded inside any larger framework for dynamic resource management in CR-LTENs which considers the channel aggregation in the MAC or PHY layer.

For a single channel with S states, let $p^i_{j,j'}$ represent the transition probability of channel number i from state j to j', which depends on the interference and fading models, and P^i denote the matrix of transition probabilities for channel i for a timeslot of duration T (generally, P^i could also be different for different users). Generally, it is not required that the CR users have prior knowledge about these probabilities. However, when this information is

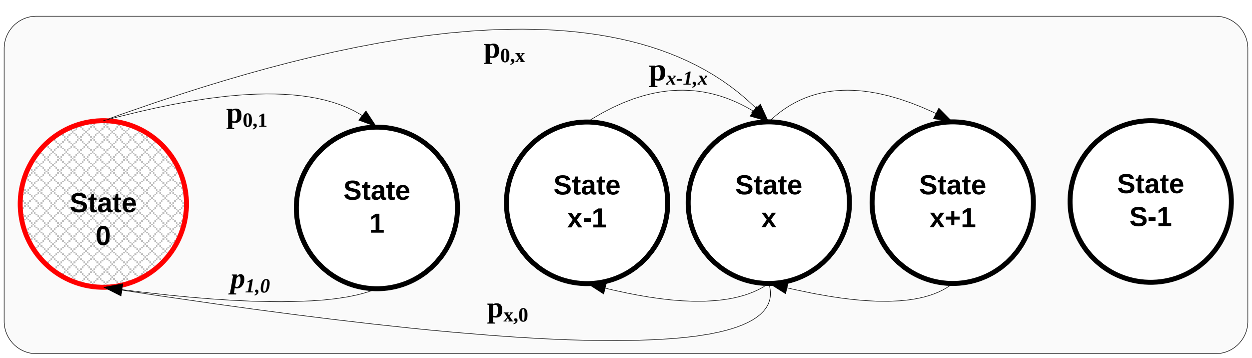

Figure 8.1 Markov chain channel model with S possible states. Sample transitions are shown.

available, the efficiency of the restoration method can be increased. Furthermore, the steady-state probabilities of the channel model Markov chain are given by π_j^i. We may drop the index i when our discussion is applicable to any channel.

8.2.2 Sensing and Channel Selection in CR-LTENs

Suppose that there are N licensed channels which can be used by the LTE network. When a UE wants to perform the restoration, it prepares (or receives from eNB) a list of channels called *channel sensing order* (CSO) to be sensed one by one and follows this order until it finds a channel with acceptable characteristics. The permutation (order) of the channels in the list depends on the policy that the node employs, called a *channel search or sensing scheme*. For example, in the random scheme, the channels are randomly sorted in the list, while in a history-aware scheme, as will be discussed shortly, the channels are sorted based on their background information. Sensing represents estimating the quality of the channel, which is done by estimating the interference level and fading coefficients. When the channel is free, pilot signals can be transmitted for such a purpose Astely *et al.* (2009). Some OFDM symbols at the beginning of each LTE frame can be used for this purpose. This implies that sensing is performed simultaneously by both ends of a link and channel selection will be done after required negotiations. Returning to the channel model, sensing is equivalent to finding the current state of the channel out of S possible states.

For sensing each channel in the list, a timing cost called *switching time (T_w)* to align the radio from one channel to the next channel is added to the *sensing time (T_s)*, spent for checking the status of the channel. Assume that the sensing and switching time are approximately the same for all channels (otherwise, we can consider the average of them, which is a realistic assumption when the channel bandwidths are similar). With this assumption, the time spent for verifying each channel is given by $(T_s + T_w)$, which we call a restoration *minislot*. When the CR node makes the decision to select a channel after sensing X channels (not accounting the previous operating channel of the user, which is sensed first before performing the restoration), the restoration time is given by $T_r = X * (T_s + T_w)$. X is a random variable

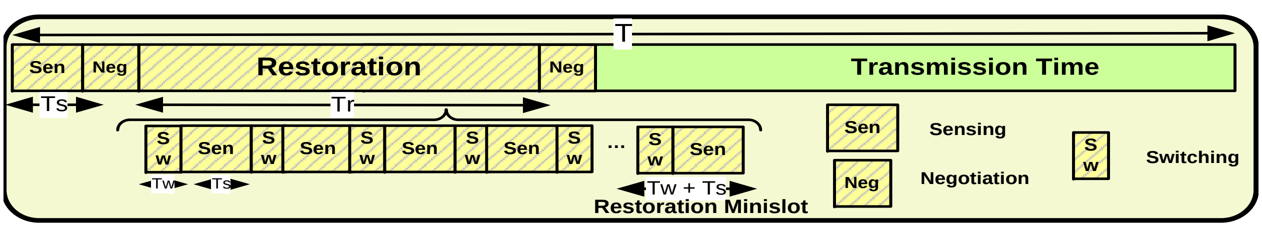

Figure 8.2 Timeslot structure and the structure of the restoration time.

which depends on the restoration scheme.

If we assume that the duration of the timeslots (LTE frames) is equal to T, as illustrated in Figure 8.2, the throughput in each timeslot is a function of the restoration time T_r and the quality (state) of the found channel. All timing parameters then can be defined based on the number of OFDM symbols, which can be transmitted. We define the throughput as the average number of fixed-size packets that a user transmits per timeslot in the remaining portion of the timeslot after the restoration. Furthermore, the maximum nominal number of packets that can be transmitted over a channel with unit bandwidth is equal to T when the channel is in its best state (Z_{S-1}). The throughput of the user when it starts operating over a channel in state j, when this channel has been selected after a restoration with a timing cost equal to T_r, can be thus given by :

$$A = (T - T_s - T_r) * B_j, \tag{8.1}$$

where B_j is the service capacity of the selected channel in state j. The throughput will be used as the performance metric to compare our new restoration scheme with other restoration schemes.

8.3 History-awareness and Greedy Restoration Schemes

In this section, we first review the proposed framework for history-aware channel selection schemes in CR-based networks Azarfar *et al.* (2011b). Afterwards, our greedy restoration scheme is presented, which is an extension of the idea proposed in Azarfar *et al.* (2011a) to a general network with heterogeneous channels. At the end, the unification of those approaches is discussed.

8.3.1 History-aware Channel Search Schemes

The basic cognitive features assumed for a CR-based network are learning capability and history-awareness Mitola III (2000); Clancy *et al.* (2007). For channel searching methods, this capability can be employed to compute a historical metric for each channel in the form of a *quality score (QS)*, where the QS of a channel is a metric representing the desirability of the channel to the users. During the restoration, a CR node can use this quality score (QS) to reduce the decision space, which increases the chance of success and the quality of the result in a shorter time. Such algorithms are called history-aware (HA) search schemes.

The QS may be defined differently in different networks. For instance, in the simplest case, the channel availability can be considered as a binary (0/1) quality score. In this paper, as our objective is maximizing the throughput, the QS is defined based on the service capacity of each channel such that the current state j of channel i, $Z^i = j, j = 0, 1, ..., S-1$, defines Q^i_j, the instantaneous quality score of the channel number i. Q^i_j is the primary quality parameter of a channel as indicated in Table 8.1. It is assumed that the channel desirability to a user increases with Q^i_j.

The second quality parameter of a channel is q^i, the average or steady-state value of the channel i quality. Finally, we define a quality index I^i which is the parameter used in order to sort the channels to create a sorted channel sensing order (CSO). In the simplest case, I^i is equal to q^i :

$$I^i(n) = q^i(n) \tag{8.2}$$

but in general, it can be a function of q^i and other parameters such as the variance of q^i.

We propose in our framework that each user keep a Spectrum Information Table (SIT) as illustrated in Fig. 8.3 to create the quality index. In a centralized approach, only the central entity may keep and update the SIT. However, in general, since the location and spectrum sensing results of distinct users could be different, the SIT of two different users could be dissimilar in some instants of time.

First, the user needs to update q^i based on the Q^i_j obtained from direct channel sensing or through new information about the state of a channel received from a central entity. As an example, q^i can be updated using the following moving average :

$$q^i(n) = \gamma Q^i_j(n) + (1-\gamma) q^i(n-1) \tag{8.3}$$

where n, j and γ respectively represent the current timeslot (time index), the current state of the channel and the parameter of the moving average. When there is no update for a channel, the user keeps the same quality score (i.e., $q^i(n) = q^i(n-1)$).

The user also updates l^i, the last sensed/received state of the channel, based on the current sensed state. To insure up-to-date information, there is a counter (a^i) that represents the age of the QS for each channel that indicates when q^i and l^i have been updated for the last time. When QS is updated, this counter is reset to zero. Then, in each timeslot the counter increases by one. If the CR does not sense this channel again before the specified

Table 8.1 Quality parameters. Time index (n) is dropped.

Notation	**Description**
Q^i_j	Instantaneous quality of channel i in state j
q^i	Long-term quality of channel i
I^i	Quality Index, a function for sorting the channels

↓ Parameters Channels →	1	2	...	N
Quality Score (QS)	q^1	q^2		q^N
Validity Flag (VF)	v^1	v^2		v^N
Counter/Age (Cnt)	a^1	a^2		a^N
Last Sensed State (LS)	l^1	l^2		l^N

Figure 8.3 Spectrum Information Table (SIT).

timing threshold or does not receive any status update, the counter reaches its maximum and the Validity Flag (VF) (v^i) is raised to indicate that q^i is out-of-date and thus invalid. This framework is general and is independent of the MAC protocol. It can therefore be added as a new feature to any CR-based network, such as LTE networks. For instance, the Spectrum Information Table (SIT) can be kept in each eNB to keep the UE simple and cost-effective. When channel estimation is performed in each frame or sub-frame, the result, beside the action applied (link adaptation), is kept in the SIT to be used later for recovery.

As assumed so far, the network nodes generally do not know the transition probabilities of the channels. Now, if we assume that they are aware of these probabilities, e.g. through a central entity, or estimation as discussed in Haykin (2005), a HA search scheme may be further improved when the recent states of the channels are also taken into account for the decision making. In other words, the CR users also consider possible dependencies between the current state of the channel and its last sensed state. For instance, when a channel is sensed in state zero in the previous timeslot, if the transition probabilities show that the probability that this channel returns to a good state in the current timeslot is low, the user may postpone sensing this channel even if its long-term QS is high. As illustrated in Figure 8.4.b, the estimated service capacity of the channel can be used as the quality index :

$$I^i(n) = E[Q_j^i(n) | Z^i(n - a^i) = l^i] \tag{8.4}$$

where l^i stands for the state of the channel in the last timeslot, $n - a^i$, where this channel was sensed (the field LS in the SIT) and j represents the current state of the channel. As the transition probability matrix P^i is known, the expected state of channel i in this timeslot can be obtained from $(P^i)^{a^i}$, where a^i is the value of the age counter in the SIT, and the last measured QS value l^i.

At the beginning of the restoration, each UE locally sorts the channels based on $I^i(n)$ to form the CSO for this restoration period. As illustrated in Figure 8.4.a, all channels are first separated into two lists based on their validity status and then sorted based on their $I^i(n)$.

The concatenation of two lists creates the CSO for this restoration period. Note that when some channels have the same quality score, one of them may be preferred if other parameters, such as switching time, are taken into account. Otherwise, a random selection among similar channels is performed. The user starts sensing the channels of the CSO sequentially and updates their QS according to the sensing results. Which channel will be accepted and selected is discussed in the next section.

8.3.2 Greedy Restoration Scheme

Considering the channel model and previous discussion about the cases where a user may perform the restoration, in the literature it is mostly assumed that a CR node performs the restoration either when its channel is not available anymore (state zero) or the channel quality is not acceptable (a threshold service capacity). Otherwise, it continues using the same channel to transmit a flow of packets during a call, on the channel assigned to it (for instance, by employing adaptive modulation and coding (AMC) to keep the BER constant).

In this paper, we propose a novel greedy restoration mechanism in order to provide a higher throughput for the CR-LTEN. The basis for our restoration scheme is that a user always prefers to have a better channel and thus may perform the restoration periodically (e.g., at the beginning of each timeslot in a time-slotted network model) even if its channel is still available. However, the time spent for restoration should be proportional to the current channel state. If the CR's current channel is in state zero, which means the channel cannot be used anymore, the user may spend the longest time to find a new channel, and when its current channel is a wide enough channel in state $S-1$, the user does not perform any restoration.

Formally, the scheme is defined as follows : At the beginning of each timeslot (i.e., periodic restoration periods), the UE first senses its current operating channel. The current service

```
for   i=1:N
        I^i(n)=q^i(n)
        if   v^i==0
                add_to_list(VChs,i)
        else
                add_to_list(IChs,i)
        end
end
SVChs = sort (VChs,I^i(n))
SIChs = sort (IChs,I^i(n))
Sensing_Order (n) = concat (SVChs, SIChs)
```

a) No knowledge about the channels

```
for   i=1:N
        I^i(n)= E[Q^i(n)|Z^i(n-a^i)=l^i]
        if   v^i==0
                add_to_list(VChs,i)
        else
                add_to_list(IChs,i)
        end
end
SVChs = sort (VChs,I^i(n),q^i(n))*
SIChs = sort (IChs, q^i(n))
Sensing_Order (n) = concat (SVChs, SIChs)

*[ if I^i(n)== I^j(n) sort based on q^i(n) and q^j(n)]
```

b) Transition probs. are known

Figure 8.4 Pseudo code for sorting the channels.

capacity of the channel determines the maximum length of the restoration duration, which is the maximum number of channels to be sensed before making the decision. Suppose that L ($L \leq N$) represents the maximum number of channels that can be sensed in one timeslot. Then, for a current channel i in state j, we want to find a parameter called $k_j^i \leq L$ which is the optimal maximum length of the restoration and is a function of several parameters such as the number of channels and users and the service capacity of the channels in different states. The user senses at most these k_j^i channels and then selects the best channel among them, which implies that *channel recall* is possible as assumed in other papers, e.g., Chang et Liu (2007). During recovery, the user may sense one channel usable, but continues the recovery hoping to find a better channel. When the user returns and decides to use a channel sensed before, it is called *channel recall*. The sensing time is generally much shorter than the time of channel variations, so the probability that when the user returns to a previously sensed channel, the channel state has changed is negligible.

An additional aspect is that the decision of the number of channels to be sensed, k_j^i, is not static and is updated after each minislot (sensing each channel) based on the sensing results during the restoration. We thus define an updating function $F_u(.)$ that receives the sensing results up to now in this sensing period and returns the optimal maximum number of channels to be sensed from now on. When this function returns zero, the recovery is ended and the user selects the best channel already sensed. Let us assume that the bandwidth of all channels is the same such that the service capacity of the channels can be represented only by their states (homogeneous channels). We discuss shortly after the more general heterogeneous case. With this assumption, the updating function takes the best state (service capacity) that has been found so far and the number of channels previously sensed. For instance, if a user finds a channel in the best state (S-1), the restoration is ended and the user starts using this channel. In other words, as the length of the restoration is updated in each minislot, if the output of the updating function is not zero, the user continues the restoration for one more minislot and then calls the function again. Otherwise, the user ends the restoration. For further examples, please see Figure 5 in Azarfar *et al.* (2011a).

To be able to find the optimal value of the restoration duration for a typical network, let us first assume that there are N homogeneous channels with the same occupancy distribution, bandwidth and transition probabilities, which is a realistic assumption in several traditional wireless networks. Moreover, we focus on a single link where one user is communicating with an eNB over N opportunistic channels. This implies that N does not necessarily represent all the available channels in the network. We define two new notations : N_j represents the average number of channels which are in state j (in each restoration period) and can be given by :

$$N_j = N\pi_j \tag{8.5}$$

and N_j^c is the cumulative average number of channels in a state with a state number less than or equal to j. It can be written as :

$$N_j^c = N \sum_{g=0}^{j} \pi_g. \tag{8.6}$$

The updating function finds the best number of channels to be sensed by maximizing the average estimated throughput of the user. Suppose that the best channel that the user has found so far is in state j and the number of sensed channels so far is represented by a variable h. If the restoration continues for k other minislots ($0 \leq k \leq L-h$), we can write the expected value of the estimated throughput as :

$$E[A] = (T - hT_s - kT_s)\left[B_{j+1}\theta_{j+1} + ... + B_{S-1}\theta_{S-1} + B_j(1 - \sum_{l=j+1}^{S-1} \theta_l)\right] \tag{8.7}$$

where the last Σ stands for the case that the user cannot find a better channel and thus stays in (returns to) the best found channel in state j (all the k channels are in a state lower than or equal to j). Without loss of generality, we have ignored the switching time, T_w. θ_x is the probability that the best channel that the user finds among these k channels is a channel in state x ($x > j$). θ_x can be written as :

$$\theta_x = \begin{cases} \frac{P(N_x^c - h, k) - P(N_{x-1}^c - h, k)}{P(N-h,k)}, & x > j \\ 0, & \text{Otherwise.} \end{cases} \tag{8.8}$$

where $P(n, k)$ represents the permutation of (n, k), equal to $\frac{n!}{(n-k)!}$. Note that, for the permutation function, the inputs should be integer numbers, but this is not necessarily the case in this work. Thus, a rounding function is used which results in some rounding errors. By maximizing Eq. (8.7) over k, one can find the optimal k^* which maximizes the throughput.

$$F_u(.) = arg \max_k \{E[A]\}. \tag{8.9}$$

When the returned optimal k^* is zero, the recovery is terminated. The pseudo-code of the restoration scheme is shown in Figure 8.5.

So far we have discussed the proposed restoration scheme for homogeneous channels ; however, it is applicable to any network model where the channels may have different bandwidth

```
%% Ch*=Current operating channel
%% BS=Best state has been found so far
%% k=Optimal number of channels to be sensed
%% Chn=Current channel being sensed in the sensing order
%% h=Number of sensed channels
h=1;
BS=state(Ch*);
k=Fu(BS,1);
while k>0 && h<N
    k=k-1;
    h=h+1;
    if state(Chn) > BS
        SelectedChan=Chn;
        BS=state(Chn);
    end
    k=Fu(BS,h);
    if k==0 return SelectedChan;
end
```

Figure 8.5 Pseudo-code of the proposed restoration scheme. $Fu(.)$ stands for the updating function.

and dissimilar transition probabilities, which is the case in LTE networks. The only requirement to find the stopping time in the proposed restoration scheme is to have knowledge of the channels' bandwidth and steady-state probabilities, which is a realistic assumption in any LTE network in operation. In this case, the service capacity of the state j of channel i with bandwidth W^i is given by B^i_j. The probability to find channel i in state j during the restoration can be given by $\frac{\pi^i_j}{N-h}$, if we assume that the scheme decides uniformly between the remaining not-sensed channels.

When a user is operating over channel i in state j, the probability of finding a better channel in the next minislots of the restoration can be calculated from this available knowledge about the steady states of the channels. For this goal, when a channel is sensed, it is eliminated from the decision space and new probabilities for the remaining channels are calculated, i.e., the probabilities are dynamically updated.

Let's sort all possible values of service capacity in a vector called C. Then, $C(0)$ is the minimum possible service capacity which is equal to zero (e.g., any channel in state zero) and $C(Length(C)-1)$ is the maximum service capacity which belongs to the channel(s) with the largest bandwidth in the best state $(S-1)$. We define a new binary variable $u^i_{j,k}$ as follows :

$$u^i_{j,k} = \begin{cases} 1, & \text{If } B^i_j = C(k) \\ 0, & \text{Otherwise.} \end{cases} \tag{8.10}$$

Using this variable, θ_x, the probability to find a channel with a service capacity equal to $C(x)$, can be given by :

$$\theta_x = Pr\left(\text{Found Service Capacity} = C(x)\right) = \frac{1}{N-h}\sum_{i=1}^{N} y^i \sum_{j=0}^{S-1} u^i_{j,x}\pi^i_j \tag{8.11}$$

where y^i is another binary variable which represents whether channel i has been already sensed or not, and h still stands for the number of channels sensed so far (i.e., $h = \sum_{i=1}^{N}(1-y^i)$). If channel i belongs to the remaining channels not sensed yet then $y^i = 1$, otherwise $y^i = 0$.

Then in each restoration step, the average throughput that can be obtained if the restoration continues for one more minislot is given by :

$$E[A] = (T - hT_s - T_s)\left[\sum_{l=j+1}^{Length(C)-1} \theta_l C(l) + C(j)(1 - \sum_{l=j+1}^{Length(C)-1} \theta_l)\right] \tag{8.12}$$

where it is assumed that the best found service capacity of the user is equal to $C(j)$. If the estimated throughput is larger than the one that can be obtained by ending the restoration in this minislot, which is equal to $(T - hT_s)C(j)$, the restoration continues for one more minislot. Otherwise, it is ended. In other words, $F_u(.)$ returns zero if $E[A] \leq (T - hT_s)C(j)$, which terminates the recovery and the user selects the channel with service capacity $C(j)$ as its new channel.

In this equation, we assumed that even for channels with different bandwidth, the sensing time can be estimated at the same value. Otherwise, the T_s should be substituted by the expected value of the sensing time of different channels (i.e., $\bar{T_s}$).

8.3.3 History-aware Greedy Restoration Scheme

Up to now, we described a greedy restoration scheme which can be employed with any channel search scheme. That is, the order of channels to be sensed during the recovery is selected randomly. For a history-aware (HA) greedy scheme, the history-awareness is used to compute the channel sensing order employed by the greedy scheme. The role of the history-awareness is twofold : First, the CSO is created based on the background of the channels as discussed earlier (Eq. (8.2) or (8.4)) ; second, when a channel is sensed during the restoration, its QS in the SIT is updated for future restorations using Eq. (8.3).

This implies that we can list four different types of restoration schemes which are the combination of HA versus non-HA (e.g., random or sequential search schemes) and greedy versus threshold-based restoration schemes. Threshold-based schemes use a fixed value of service capacity (a fixed state when channels are homogeneous) to trigger the restoration Azarfar *et al.* (2011a). These schemes continue using the same channel as long as the service capacity of the channel remains greater than the predefined threshold, and during the restoration, they select the first channel whose service capacity is larger than the threshold.

8.4 LTE-Advanced Framework

The proposed history-aware greedy restoration scheme is a general concept applicable to different communication systems. However, for each specific system, more investigation and implementation details are still required. In this section, we specifically discuss how the proposed scheme could be integrated in LTE cognitive networks. Although this proposal shows the general guidelines of how to integrate this framework into LTE cognitive networks, implementation details still require some investigation and will be the topic of further research.

We assume in this paper a time-slotted system where the timeslot can be an LTE frame of 10 ms (it can also be any multiple of a frame). At the beginning of the timeslot, the LTE eNB can use the physical downlink control channel (PDCCH) of the first subframe to instruct one, multiple or all UEs to sense in the current subframe one or more channels. If the eNB is equipped with multiple transceivers it can use a subset of transceivers to probe the other channels. Otherwise, no data transmission is scheduled in the current subframe. The eNBs then transmit on the sensed channel during one or more OFDM symbols reference signals on selected subcarriers while the other subcarriers are left unoccupied. Since the eNB and UEs derive their clock and RF carrier from a single reference, time and frequency synchronization can be maintained on different channels. The UEs can thus estimate the downlink channel from those reference signals (if perfect synchronization can not be achieved, the UEs can estimate the receive signal strength on those subcarriers), and from the idle subcarriers, it can estimate the interference level or the presence of the primary user. The procedure is repeated for as many channels that can be sensed in a subframe. A similar procedure then takes place for the uplink channel in the next subframe. Furthermore, the UEs can use the uplink transmission to report to the eNB the channel measurements on some subcarriers. From those uplink and downlink measurements, the eNBs then derive, as a function of the physical layer parameters (MIMO capabilities, link adaptation, etc.), the aggregate data rate that could be achieved on the sensed channels. Based on those results, the eNBs then decide if the sensing procedure should be continued or stopped. In the former case, the eNB instructs, in the PDCCH of the next subframe, the UEs to sense other channels. In the latter case, the eNB indicates, in the PDCCH of the next subframe, the new channel that should be used by the UEs. If several transceivers are available, then only a subgroup of UEs could be switched to another channel depending on the sensing results. Although in this paper we considered a single user case, the results can be easily extended to multiple users based on the aggregated data rate. However, the multiple transceiver case still requires further investigation.

From the channel aggregation point of view, in-band channel aggregation can be easily taken into account in the proposed scheme because the overall capacity of some available channels in the same band can be seen as a wider logical channel. Wideband sensing will play the main role to provide the possibility of sensing multiple channels at the same time, during the recovery period. However, out-of-band aggregation can not be addressed very well in the proposed scheme since our scheme stops after finding an appropriate channel. We are working now on a similar approach for channel aggregation where during the recovery, the user stops when the total capacity found (aggregated throughput) is optimal.

History-awareness is mostly independent of the specific details of the PHY implementation and, similarly to other cognitive features (e.g., in Vučević *et al.* (2011); Saatsakis *et al.* (2008)), it can be implemented in the LTE base stations. The Spectrum Information Table (SIT) can be kept in each eNB to keep the UE simple and cost-effective. When channel estimation is performed, the result, beside the action applied (link adaptation), is kept in the SIT to be used for future restorations.

Finally, it is worth noting that in an LTE network with known primary users, such as TV channels, the assumption that the number of channels, their bandwidth and the distribution of channels' occupancy are known is also realistic. This knowledge is either known in advance based on the regulatory information, or can be obtained statistically after an operation period.

8.5 Simulation Results

To evaluate the performance of the proposed HA greedy restoration scheme for different channel cases, the transition probabilities of the channel model, introduced in section 8.2.1, are selected randomly between 0.05 and 0.5. The number of channels N is equal to 10 and 18 (only for homogeneous channels). L is assumed to be equal to N. The number of states, S, is selected equal to 7. The normalized service capacity of a channel over bandwidth is divided linearly among the states such that for a 7-state channel, it can be represented as $(0, 0.17, 0.33, 0.5, 0.67, 0.83, 1)$. The CSO formation is based on Eq. (8.4) assuming known transition probabilities. The parameter of the moving average, γ, is selected equal to 0.1.

The time required to handshake and confirm the change in the operating channel and the switching time (T_w) are assumed negligible. The length of the timeslot, T, is assumed equal to 10ms Astely *et al.* (2009). The experiments have been done for two values of sensing times : 500us and 900us. A Monte Carlo simulation has been carried out using MATLAB ; due to the time limit, the same experiment has only been repeated 200 times (for homogeneous channels) and 500 times (for heterogeneous channels). In each experiment for the homogeneous case, a new channel with random transition probabilities is generated and for heterogeneous case, N new channels with new random transition probabilities and bandwidth are generated. The simulation in each run lasts for 70000 timeslots.

8.5.1 Homogeneous Channels

For the case with homogeneous channels, all channels have similar transition probabilities and a unit bandwidth which implies that the maximum service capacity of all channels is assumed equal to 1. The maximum throughput of the user in one timeslot is equal to $T - T_s$ packets (Eq. (8.1)) when its channel is in the state $S - 1$ and the user does not perform any restoration (T_r=0).

We compare our proposed mechanism called "'HA-Gr"' with three other schemes : a pure greedy scheme (Gr) as discussed earlier without history-awareness, a history-aware threshold-based scheme (no-greediness) which uses the state zero as the threshold state to trigger the restoration (HA-St0) and finally another threshold-based HA scheme (no-greediness) which uses the state three of the channel as the threshold state (HA-St3) (Azarfar *et al.* (2011a)).

In Figure 8.6, the cumulative distribution function (CDF) of the percentage of relative throughput improvement (e.g., $\frac{A(HA-Gr)-A(Gr)}{A(Gr)}$), which can be obtained using the proposed scheme, is shown. As can be seen, our proposed scheme outperforms other schemes in any scenario ; however, compared to a pure greedy scheme, the improvement is low. As discussed in Azarfar *et al.* (2011b), the reason is the homogeneity of the channels which fades the role of history-awareness. That is, learning is more effective when the channels are dissimilar. Compared to a pure HA scheme which uses state three as threshold, the improvement is less than 5%. This result can be explained as discussed in Azarfar *et al.* (2011a). In that paper, we saw that when a middle state (i.e., a mean service capacity) is selected as the threshold, the performance of the threshold-based restoration scheme is close to a greedy scheme. Here, both schemes are equipped with history-awareness, so the same performance difference can be observed. Considering previous results and discussions above, it is expected that the throughput improvement will be higher if we compare the proposed HA greedy scheme with a non-HA scheme which uses state three as threshold (will be discussed shortly in Section 8.5.2).

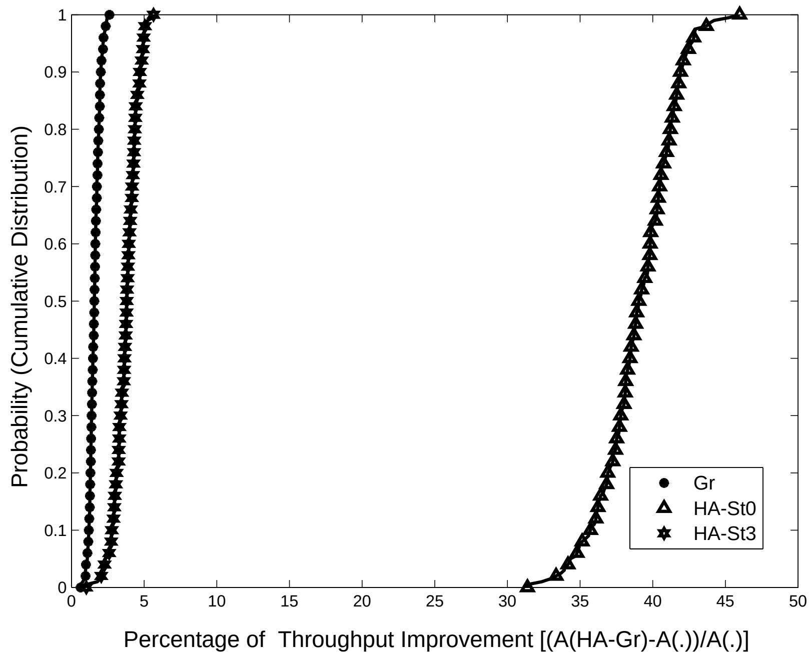

Figure 8.6 Cumulative Distribution Function (CDF) of the throughput improvement employing a HA-greedy scheme versus other schemes (homogeneous channels, N=10 and T_s =500us).

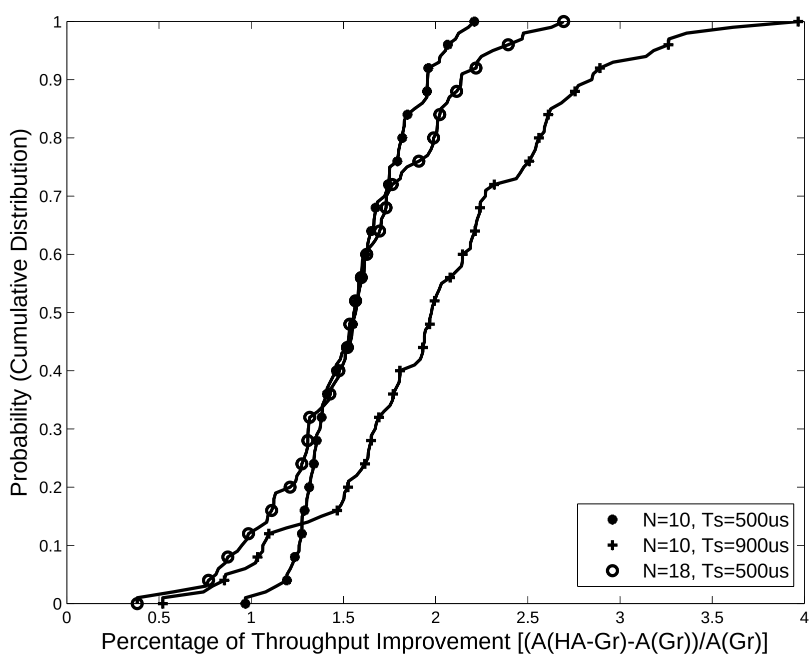

Figure 8.7 Cumulative Distribution Function (CDF) of the throughput improvement versus the pure greedy scheme (different scenarios).

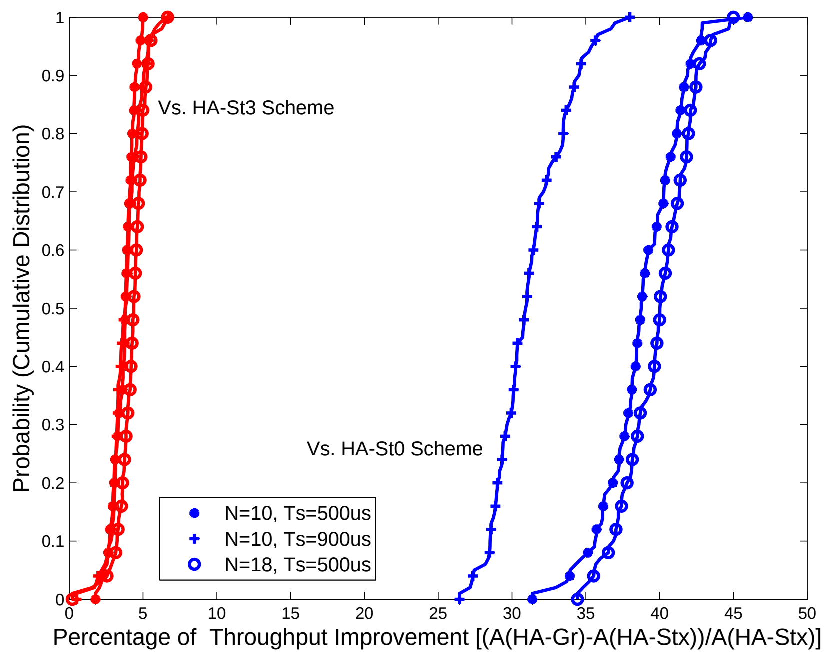

Figure 8.8 Cumulative Distribution Function (CDF) of the throughput improvement versus two threshold-based HA schemes (different scenarios).

In Figures 8.7 and 8.8, the same metrics for different values of N and T_s are compared. As can be seen in Figure 8.7, when the sensing time increases, the usefulness of a history-aware restoration scheme is more highlighted because the cost of a random restoration linearly increases with the sensing time. As history-awareness provides a sorted list of channels, the superiority of a HA greedy scheme and consequently throughput improvement versus a pure greedy scheme increases when the sensing time increases.

Compared to pure HA schemes, we can see in Figure 8.8 that the throughput improvement of the proposed scheme versus HA schemes decreases when the sensing time increases, which is contrary to the observation that we had in the comparison of two greedy schemes. The reason is that a greedy scheme explores the spectrum frequently while a threshold-based scheme, especially when a lower state is selected as the threshold, does that rarely. With the increase of the sensing time, the cost of spectrum exploration increases which decreases the superiority of a greedy scheme versus threshold-based schemes. Therefore, the amount of throughput improvement decreases.

As none of the schemes senses all the channels to select a new channel, we can see that the impact of the number of channels is not significant. However, with the increase of the number of channels, a small throughput improvement can be observed, which demonstrates that the proposed scheme can do even better when the number of channels is higher.

8.5.2 Heterogeneous Channels

In Figure 8.9, we have repeated the experiments for N=10 dissimilar channels (T_s=500us). The bandwidth of the channels are selected randomly between 1.4MHz and 20MHz. The transition probabilities are also chosen randomly where these probabilities can be different for different channels. We compare our proposed HA-greedy scheme with four other schemes : two history-aware schemes, in black, which use a threshold service capacity (not a state as for homogeneous channels) to trigger the restoration. Considering the sorted vector of all service capacities, C, the first one uses the mean of C as the threshold value (named "'HA-St-Mean"') and the second one uses the median of C as the threshold value (named "'HA-St-Median"'). "'Gr"' represents a pure greedy scheme without history-awareness and finally "'St-Mean"' stands for a threshold-based scheme which is neither greedy nor history-aware and uses the mean of C as the threshold value.

It can be seen that compared to the case with homogeneous channels, throughput improvement versus both a pure greedy scheme and a history-aware threshold-based scheme is higher. As already discussed, it comes from the fact that history-awareness is more useful and effective when the channels are dissimilar. It demonstrates that the proposed HA-greedy scheme can be effectively employed in a CR-LTEN, where the channels are generally dissimilar, to provide a significant throughput improvement.

Finally in Figure 8.10, the impact of sensing time on the performance improvement is illustrated. For a better presentation, the results are compared only versus a pure greedy and a HA threshold-based scheme. Similar to results of Figures 8.7 and 8.8, it can be seen that when the sensing time increases, the performance improvement versus a pure greedy scheme increases while compared to a pure history-aware scheme decreases.

8.6 Conclusion

Inspired by the idea of using history-aware and greedy restoration schemes in our previous research work, we proposed in this paper a history-aware greedy restoration scheme which combines the strengths of learning and history-awareness and greediness to propose a restoration scheme with the goal of increasing the overall throughput in cognitive radio-based LTE networks. The duration of the restoration scheme (number of channels to be sensed) and the stopping rule is obtained using a recursive function which takes the best channel and the number of channels sensed so far and gives the optimal stopping time over a sorted list of channels, which is prepared based on the history of the channels. Simulation results show that compared to existing restoration schemes which are only greedy or history-aware, our scheme improves the throughput in all cases. Furthermore, as the proposed scheme pro-

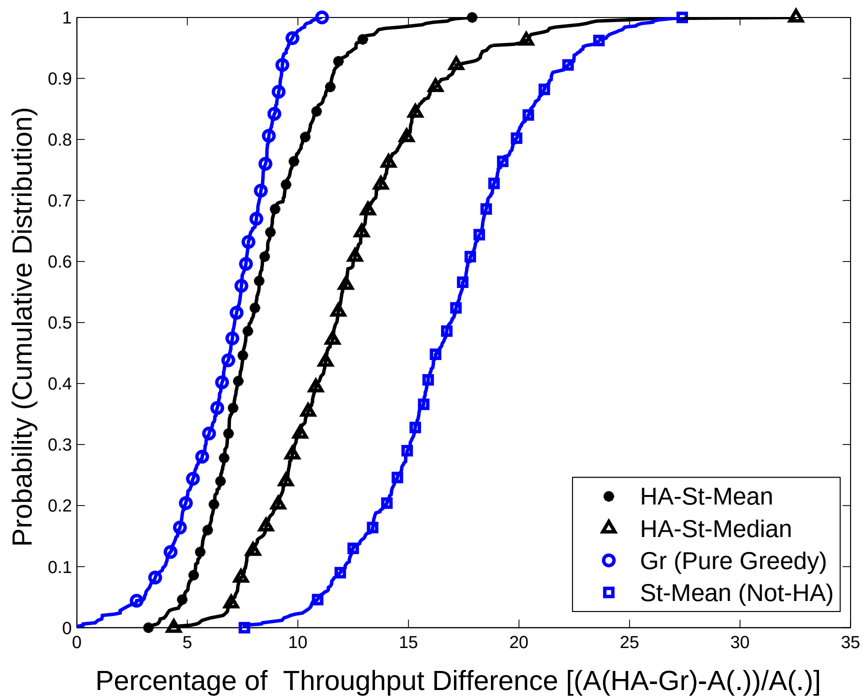

Figure 8.9 Cumulative Distribution Function (CDF) of the throughput improvement employing a HA greedy scheme versus other schemes (heterogeneous channels, N=10 and T_s =500us).

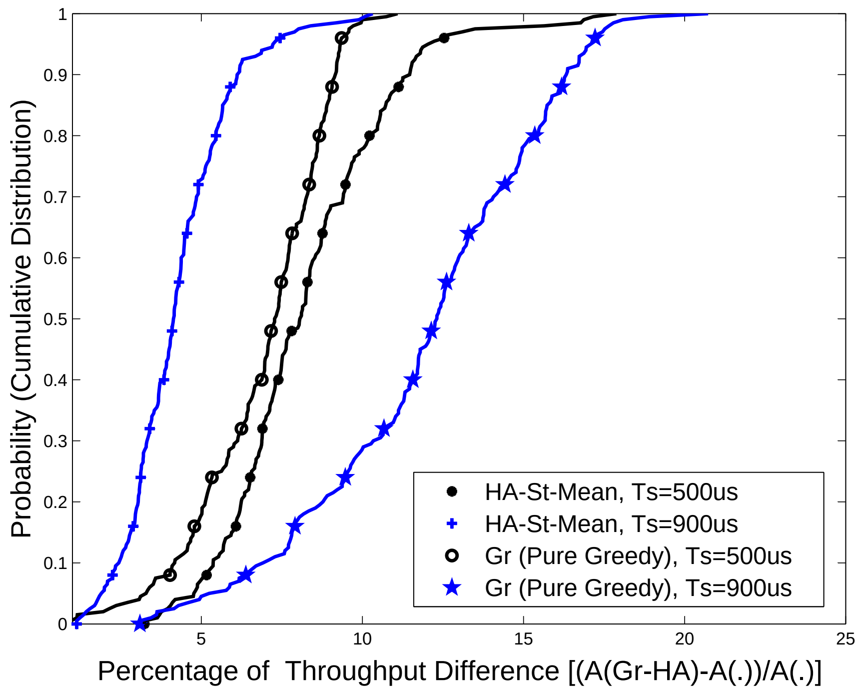

Figure 8.10 Cumulative Distribution Function (CDF) of the throughput improvement for two different values of sensing time.

vides optimal results in any scenario with no need to calculate a pre-defined threshold, it is also more rewarding compared to a history-aware threshold-based scheme which may have a similar performance.

This history-aware greedy scheme is mostly general and applicable to many communication systems such as LTE networks. It was discussed how it can be integrated into LTE networks considering MIMO and in-band channel aggregation features of LTE-advanced. The proposed scheme can be embedded inside any larger framework for dynamic resource management in LTE networks based on cognitive radios. Other requirements of LTE-advanced networks, such as out-of-band channel aggregation, will be discussed in our future work.

8.7 Further discussions (Not a part of the paper)

What has been proposed in this article chapter is a greedy restoration scheme which can be categorized in a larger framework called finding the optimal sensing order and the optimal stopping time during the recovery. Given a list of channels to be sensed, where this list may be created randomly or based on a history-aware scheme (as was the case in this chapter), the CR node starts sensing the channels one by one and in each iteration decides either to continue the recovery, or to employ a channel already found. Note that even if the sensing order has not been given, the process of optimal decision making can also determine the sensing order meaning that it selects the next optimal channel to be sensed if the decision is to continue the recovery. Dynamic programming is often used to solve the decision making process discussed above. However, due to complexity of dynamic programming, it is important to have sub-optimal heuristics. In the following, for completeness, the general dynamic programming model is discussed where the problem is studied assuming different cases : when the decision is made greedily in the beginning of each timeslot or a restoration is triggered only when the channel becomes unavailable (should be vacated) ; when channel recall is possible or when the channel recall is not applicable ; and when a single channel is selected or when multiple channels may be selected.

Assume that the probability that each channel i is available (a state greater than zero in the model discussed in this chapter) is given by θ_i. Since channels are heterogeneous, the sensing time of each could be different, so it is given by $T_{s,i}$. The parameter γ_i represents the state of the channel whether it is available or not. Current operating channel of the user is called channel zero and is naturally excluded from the sensing order because in any case, this channel is sensed in the beginning of each timeslot.

Let us start with the case where recall is not possible, so the user can not return and select a channel which has been sensed before. In the following, we propose a dynamic programming (DP) model with no recall Bertsekas (2005) where the decision variable is the channel to be selected for sensing and use. The *stages* of the model are after sensing each channel during the recovery. The *state variables*, represented in the form (S, b_k, k, γ_k), is the set of remaining channels which has not yet been sensed in this recovery period (S), and the service capacity (b_k), index (k) and the availability status (γ_k) of the last sensed channel. $b_k = 0$ represents that the last sensed channel has not been available ($\gamma_k = 0$). In the beginning of each recovery period, we have $S = \mathbb{S}$ where $\mathbb{S}$ is the set of all channels to be sensed (excluding current operating channel), and $k = 0$.

Note that the last two parameters of the state are only included for clarity. In practice, the last sensed channel is the last channel added to S, so it can be found from S, and the availability can be found from the channel quality (zero, if the channel is not available). Since the problem is discussed for a single timeslot, as was the case in the paper, the reward is given by Eq. (8.1) where T_r is the total time spent for recovery, and B_j is the service capacity of the selected channel.

The *value function* $v(S, b_k, k, \gamma_k)$ is the optimal expected reward for the given state variables. The Bellman equation Bertsekas (2005) can be given by two different equations, as follows :

$$v(S, b_k = 0, k, 0) = \max_{i \in S} \left\{ (1 - \theta_i) v(S - \{i\}, 0, i, 0) + \theta_i \mathbb{E}\big[v(S - \{i\}, B_i, i, 1) \big] \right\}, \tag{8.13}$$

$$v(S, b_k, k, 1) = \max \left\{ A(\mathbb{S} - S, T_{s,0}, b_k), v(S, 0, k, 0) \right\}, \tag{8.14}$$

where θ_i is the probability that channel i is available. The throughput (reward) $A(\mathbb{S} - S, T_{s,0}, b_k)$ is given by :

$$A(\mathbb{S} - S, T_{s,0}, b_k) = (T - T_{s,0} - T_r) b_k, \tag{8.15}$$

where T_r is given by :

$$T_r = \sum_{i \in \mathbb{S} - S} T_{s,i} + T_w. \tag{8.16}$$

The stopping time is when we have $A(\mathbb{S} - S, T_{s,0}, b_k) > v(S, 0, k, 0)$. This problem is known as *parking problem* in the literature of dynamic programming Bertsekas (2005). The model can numerically be solved by *backward induction* Bertsekas (2005). The optimal policy is a threshold policy meaning that in each iteration of the recovery, if the service capacity of the found channel is higher than a threshold value, it is optimal to stop the recovery and use the found channel. Otherwise, the recovery is continued. The threshold value is different in each iteration and is naturally decreasing, so in the last iteration any available channel is selected (the threshold value is zero). This is also the terminal condition. Depending on how channel service capacity is defined, we may be able to analytically find those threshold values. For instance, for channel capacities defined based on a continuous fading process, threshold values have been found in Jiang *et al.* (2009).

The problem discussed so far was with no channel recall. When the user is able to use a channel previously sensed during the recovery, the model will be similar, with some small modifications. First, the best channel found so far is kept as a state variable, b_{bst}. The second difference is that in both cases where the last sensed channel is or is not available, the recovery may be terminated and the user decides to use the best channel found so far. Finally, due to the switching time required to align the radio to a previously sensed channel, the user may decide to use the last found channel even if its service capacity is lower than the best channel previously found. For the simplicity of presentation, let us define

$$G(b, S) = \max_{i \in S} \left\{ (1-\theta_i) v(S - \{i\}, 0, i, 0, b) + \theta_i \mathbb{E}\big[v(S - \{i\}, B_i, i, 1, b)\big] \right\}. \tag{8.17}$$

Then, we can write :

$$v(S, b_k = 0, k, 0, b_{bst}) = \max \left\{ A(\mathbb{S} - S, T_{s,0} + T_w, b_{bst}), G(b_{bst}, S) \right\}, \tag{8.18}$$

$$v(S, b_k, k, 1, b_{bst}) = \begin{cases} \max \left\{ A(\mathbb{S} - S, T_{s,0}, b_k), G(b_k, S) \right\} & b_k \geq b_{bst} \\ \max \left\{ A(\mathbb{S} - S, T_{s,0}, b_k), A(\mathbb{S} - S, T_{s,0} + T_w, b_{bst}), G(b_{bst}, S) \right\} & b_k < b_{bst} \end{cases} \tag{8.19}$$

The terminal condition in this model is when all channels are sensed, so the user either employs the last sensed channel or the best channel previously found.

It is worth noting that when the sensing order is given (e.g., using a history-aware scheme), only the stopping time should be determined, so the decision will be between continuing the recovery and sensing the next channel, or stopping and using a channel already found. The model will be therefore simpler than the general model discussed.

If instead of a single channel several channels can be aggregated to be employed, b_{bst} will be the sum of the capacities found so far. The switching time can be neglected. The case where instead of per timeslot, new recovery is performed only when the channel becomes unavailable, and the case where the queue status of different classes of traffic is taken into account are discussed in Azarfar *et al.* (2013a).

CHAPTER 9

GENERAL DISCUSSIONS AND APPLICATIONS

The main objective of this research work was to assess the reliability and quality of service in cognitive radio networks focusing on recovery and spectrum migration, with a quick glance at differentiation. Considering the lack of a rich literature on top of which differentiated reliability is investigated, most efforts of this work were on building the foundations. In this section, we review the results and discuss how these results can be applied to existing or incoming technologies and standards.

9.1 Results analysis

Reliability or quality of service in cognitive radio networks should be regarded from two perspectives. On one side, cognitive radios are intelligent and powerful, so that they may provide new means to improve the performance in any network where they are employed. On the other side, currently the main motivation beyond the development of cognitive radio networks is to actualize the dynamic spectrum access (DSA) paradigms. DSA brings a new root of failure and interruptions which is the need to vacate the channel in case the licensee returns to its channel. In Chapter 3 the interaction of these two sides of cognitive radio was investigated to identify which approaches can be followed to improve the reliability and quality of service in cognitive radio networks. For any cause of failure in wireless networks, it was discussed which feature(s) of cognitive radios can be used to improve the reliability or quality of service. The results of that chapter is thus the foundation of this research work and any future work on reliability improvement or analysis in cognitive radio networks.

In this research work, the focus was on interruptions and triggered recoveries due to channel vacation. In Chapter 4, we saw that the time spent for recovery prevents the network from reaching the maximum availability. In that chapter, it was assumed that the recovery time is fixed and given. The assumption was realistic because usually there is a quiet period during which all users should have no communication even if they have already found/reserved their channel. To achieve a high mean time to hard failure and low mean time to repair, an available option was to increase the number of channels, so that with a high probability, a user who missed the channel can soon find a new channel. This thus highlights the importance of having an improved recovery approach which can guarantee a minimum recovery time.

Coming to Chapter 5, the proposed queueing model highlighted that the impact of the interruptions is not linear, and can not be approximated by assuming, for instance, larger packets. Moreover, even though the ratio of the channels' availability periods or operating periods versus the recovery/interruption periods was a very important and determining factor, the performance could be significantly different for two networks with short availability and interruptions periods and with long availability and recovery periods with the same ratio of $\frac{E[Y]}{E[R]}$. This thus shows the necessity to have an accurate model to be able to compare the improvement which can be achieved in higher communication layers, when a better recovery approach is suggested. Different reliability and quality of service parameters can be found from the queueing model to find how channel parameters, such as availability and unavailability periods, and the recovery algorithm parameters, such as the recovery duration, affect packet loss, delay and jitter, and also the MTTF and MTTR for hard and soft failures ; parameters which are critical to be controlled especially in multimedia communications.

In Chapter 5, it was further discussed how priority queueing helps a CR node to implement traffic differentiation. The notion of interruption raised two new priority disciplines beside the classical disciplines of preemptive and non-preemptive. This enables the CR node to more accurately control the interaction of different classes of traffic in conjunction with the interruption periods. When the CR node is equipped with four priority disciplines, using any of these disciplines can be considered as a decision variable. The node thus selects at each instant of time, based on the queue occupancy and requirements of different classes of traffic, which of those disciplines should be employed. This idea was discussed in Azarfar *et al.* (2012b). Modeling the problem as a dynamic programming model, we showed that such an optimal decision can improve the performance (an aggregated cost as a function of the queue occupancy of different classes of traffic) compared to a random decision or always using the same discipline.

The queueing model was first studied for homogenous channels ; in other words, the service rate of the queue server remains the same. In reality, this is not the case in communication networks because even the same channel may provide a time-variant service rate due to fading or interference. The article in Chapter 6 was thus an effort to address channel heterogeneity. Based on the recovery algorithm, it was assumed that the probability of finding a channel with a given rate is known, the occupancy was thus modeled as a multi-row Markov chain. Each row represents one of the possible service rates. In scenarios with a few possible service rates, the Markov chain can numerically be solved to find the performance of the CR node. Two analytical approximations were also proposed based on the symmetrical structure of the multi-row Markov chain. Even though the multi-row Markov chain can be used only when the operating periods are exponentially distributed, we also saw that assuming an exponential distribution is not far from reality, so that the two proposed analytical approximations can be acceptable even if the operating periods are not ideally exponential. We have also observed that if the network designer has control over the transition probabilities between different channels (finding a specific service rate after an interruption), there are optimal points where the performance (especially a combination of the delay and jitter) is maximized. Transition probabilities have a random nature (e.g., due to fading), so it is not straightforward to have specific transition probabilities. However the recovery algorithm and the type and number of channels are parameters which help the designer approach specific optimal transition probabilities. For instance, if an optimal value for q_1 (probability of finding a channel of type 1) is equal to 0.5 and a random recovery results in $q_1 < 0.5$, either more channels of type 1 may be provided, or the recovery algorithm should be trained to have a higher probability of finding a channel of type 1. History-awareness can help, for instance, to have such an improvement for the recovery algorithm.

Queueing models were first provided for a single link. In multi-user scenarios especially with a random medium access protocol, there are some periods for competition and channel reservation. Based on the network model, either those periods are reiterated for each packet, or a bunch of packets are served before a new competition. In the former case, those periods are considered as a part of the service time of each packet, so the same queueing model with interruptions proposed in Chapter 5 with a modified service time can be used to analyze multi-user scenarios. In the latter scenarios, those periods are a new source of interruption, so the general queue model with the original service time, but a modified distribution for interruption periods, is employed. In the latter model, the interruptions are not only due to primary users, but also due to competition and channel reservation periods.

From the multi-user scenario with discrete-time distributions investigated in Chapter 7, we have learned that the key factor which affects the performance is the medium access protocol and access to control channel. It was observed that an Aloha-type access to control channel is a serious bottleneck such that when there are several available and idle channels, the users may compete for several timeslots until one of them succeeds in reserving a channel. This behavior is not particular to cognitive radio networks and also exists in traditional wireless networks with a random medium access. What makes this impairment factor more crucial in cognitive radio networks is that because of the frequent interruptions occurring due to the arrival of primary users, a user may visit more frequently the control channel to reserve a new data channel after an interruption. This thus explains why with an Aloha-type medium access, the performance can be better when instead of a switching policy ; i.e., vacating the channel in case of appearance of primary users, a waiting and buffering policy is employed, so that the CR user waits for the primary user to vacate the channel. Further, in addition to having efficient sensing and recovery schemes, developing efficient medium access protocols and control channel consideration in multi-user cognitive radio networks are critical.

Finally, in Chapter 8, we saw how learning from previous sensing results help a CR node improve the recovery performance. Instead of a random list of channels to be sensed, available information about the channels provide a ranking for the channels, so that the sensing order will be a sorted list of channels. The idea is general and applicable to any network model especially if spectrum sensing is frequent. For differentiation purposes, the same proposed spectrum information table will be used, but the function which sorts the channels will be different in favor of different classes of traffic. For instance, for a class of traffic, a higher bandwidth and channel rate can be more favorable, but for another class of traffic, channel stability is more important. When the first class has a higher priority, the sorting function may give a higher weight to channel service rate. Otherwise, channel stability may play a more significant role in the sorting function.

In the same chapter, a greedy restoration scheme was proposed which is a heuristic solution for the general problem of finding the optimal sensing order and recovery stopping time in cognitive radio networks. The general problem can be discussed in several ways and for different cases. For almost all possibilities, the problem can be modeled with a dynamic programming model. DP can be solved numerically, but based on the type of the problem (finite horizon versus infinite horizon, type of cost function, type of state variables, etc), an analytical solution may also be found. To address differentiation, as investigated in Azarfar *et al.* (2013a), the cost function will be a weighted function of the performance of different classes of traffic, such as queue occupancy, delay or packet loss. The stopping time is naturally when stopping the recovery and using a found channel results in a lower cost than continuing the recovery.

Considering the well-known weaknesses of dynamic programming, such as high running time and exponential increase of the possible states when a new state variable should be considered, numerically solving the DP is only applicable in small scenarios. Efficient heuristics are thus necessary for practical implementations, which was the motivation behind the greedy restoration scheme proposed in Chapter 8. The other approach to tackle the impracticability of the dynamic programming model was to try decreasing the state space. For this aim, we suggested a hierarchical recovery scheme in Azarfar *et al.* (2014b) where the channels are classified into multiple sets based on their characteristics, such as availability and service capacity. To perform the recovery, instead of working on a flat list of all channels, the cognitive radio node first selects a channel set (type) and then performs a channel migration in the selected set to search for an available channel. The focus of this paper is on the first part of the hierarchical channel handover, that is, selecting the appropriate channel set. The DP model will be much smaller because the decision variable instead of the channels is the channel sets. Recovery in the selected set can be any of the proposed schemes in the literature for a flat channel handover ; for instance, a random Luo et Roy (2007), a history-aware Azarfar *et al.* (2014c) or a DP-based scheme Jiang *et al.* (2009); Azarfar *et al.* (2013a). Not only the hierarchical model is easier to be implemented and more time-efficient to be solved, but also we observed a significant improvement compared to a random flat recovery.

9.2 Applications and practicality

After more than ten years of research and standardization efforts in the area of dynamic spectrum access and cognitive radio networks, different protocols and standards, such as IEEE 802.22, IEEE 802.11af (a.k.a, cognitive WiFi) and IEEE 802.16h (a.k.a, cognitive Wimax) are still in their prototyping and evaluation period. Among them, two protocols are closer to being implemented and widely deployed : IEEE 802.22 for wireless regional area networks (WRAN), and IEEE 802.11af known as TV white space (TVWS) WiFi. There are also some suggestions to deploy cellular LTE networks on white spaces. In the following, we briefly review those protocols and discuss how the results of this research work can be applied to those standards.

9.2.1 IEEE 802.11af

The IEEE 802.11af standard is proposed for an infrastructure network where a geolocation database-dependent (GDD) enabling station, equivalent to an access point in traditional WiFi, is responsible for spectrum management within its basic service set (BSS), as illustrated in Figure 9.1. The information on potential channels and white spaces is received from a central database called registered location secure server (RLSS), so spectrum sensing is not a required feature neither for the GDD-enabling station (AP) nor for the GDD-dependent stations (STAs). However, the proposed failure model in this thesis, due to channel vacation, still exists because the central database may inform a GDD-enabling station to exclude a channel from the list of available channels. If the network is operating on that channel, a spectrum migration is then required. Therefore, all discussions on the failure and recovery models proposed in Chapter 3 will be applicable to a TVWS BSS. Considering a single GDD-dependent station, it operates over the network-wide channel to transmit some packets and then some time is spent for competition with other users. This introduces a source of interruption. As discussed, there may be a spectrum handover in the whole network, which introduces another source of interruption, which however occurs rarely. The queuing model with interruptions can thus be employed here to find accurately the performance metrics. However, given that spectrum migrations happen infrequently, existing results in

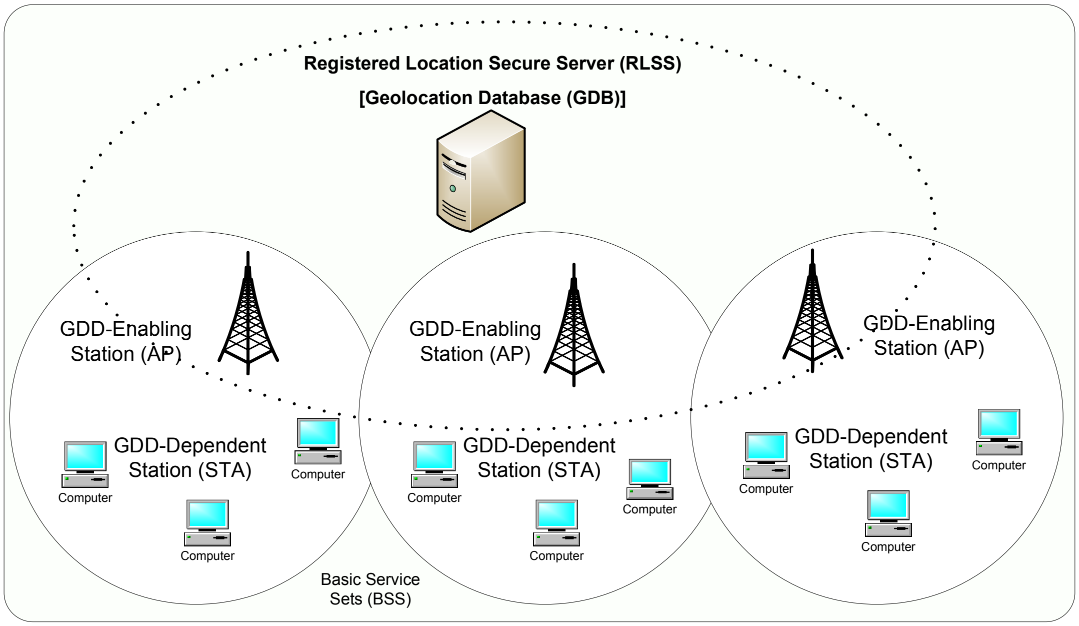

Figure 9.1 A sample IEEE 802.11af network with three BSSs served by a registered location secure server (RLSS).

the literature for IEEE 802.11 performance assessment (e.g., Bianchi (2000)) can also be employed with an acceptable accuracy. For traffic differentiation, priority queueing disciplines suggested in Chapter 5 can be used.

There is no sensing order in IEEE 802.11af networks. However, an interesting research work is to investigate how among the list of available channels received from the central database, a single channel is selected for the whole network. A channel estimation period is suggested during which the access point transmits pilot signals and the stations estimate the channels. Estimation results are then sent to the GDD-enabling access point. The AP then decides the best channel to be employed to satisfy some constraints, considering different reported channel qualities by stations.

This mechanism is similar to an election and preference aggregation where clients are the voters and channels are the candidates. Each client ranks all candidates and assign them a score, which is the measured quality. The AP should then aggregate all these preference lists to decide the final rank which is a single choice (single channel) or a final preference list (e.g., the second channel of the list can be a backup channel). As the problem is modeled as an election, the results of social choice theory Sen (1986) in addition to optimization models can be used to discuss different scenarios.

Differentiation in IEEE 802.11af networks can be applied in two parts. During the normal operation of the network, there is no significant difference between a TVWS WiFi and traditional WiFi, so existing ideas in the literature for differentiated services in IEEE 802.11, such as QoS supports in IEEE 802.11e Choi *et al.* (2003), can be used. The novel part is differentiation in the channel handover process. During the channel selection process, as discussed above, differentiation can be included in the voting and preference aggregation process. In other words, the voters will have different weights and there will be a weighted aggregation Nisan *et al.* (2007).

It is worth reminding that the history-aware recovery scheme presented in Chapter 9 can also be employed in the GDD-enabling station such that the new operating channel among available channels is also selected based on the background information available in the local database of the access point.

Considering the coexistence of several BSSs in an area, where each one operates on a single channel, there is a similar problem in a hierarchical perspective where the central database server may send different channel lists to different basic service sets. From the channel assignment point of view, it is equivalent to assume a virtual infrastructure network composed of GDD-enabling stations and the central database, so similar discussions can be provided for this virtual network.

9.2.2 IEEE 802.22

IEEE 802.22 is a standard for wireless regional area networks which targets low-population and rural areas. Similar to the previous 802.11af standard, it operates on TV white spaces in an infrastructure manner. Base stations will be equipped with geographical positioning system (GPS) and communicate with a central spectrum database to obtain spectrum availability information on the region based on their location. Most of the discussions provided for the IEEE 802.11af are therefore applicable here.

One of the differences between the two standards is that IEEE 802.22 uses OFDMA as the technology of choice in both uplink and downlink. During the normal operation of the network, differentiation is thus supported by a rich literature of schemes and suggestions on resource allocation in OFDMA networks, e.g., Han *et al.* (2005), which discussed how subcarriers should be assigned to different customer-premises equipments (CPE) in order to satisfy different constraints (per CPE power, total power, minimum bandwidth, etc) and different requirements (priority among CPEs, etc). What is different for the CR network is again the channel selection process and spectrum sensing.

In IEEE 802.22, it is strongly recommended that the base station and CPEs are also equipped with spectrum sensing capabilities. In a periodic manner or occasionally by a request from the base station, each CPE senses a list of channels and reports the results to the base station. The base station then behaves as a fusion center and combines all reports to decide if a spectrum handover is required or if any other actions should be taken. The operation of the user can be modeled with a queueing model with occasionally fixed interruption periods (about 25ms Cordeiro *et al.* (2006)) for fine sensing. Such interruptions are more frequent than spectrum decision interruptions in IEEE 802.11af, so a performance model which explicitly addresses interruptions is recommended. Note that we neglected periodic fast sensing periods because their duration is short (1ms Cordeiro *et al.* (2006)).

Further, since channel aggregation is supported in IEEE 802.22, another important decision to be made by the base station is whether to aggregate two or three channels or not. This decision depends naturally on the requirements of the CPEs. There is a tradeoff because on one side a wider channel is more preferable, but on the other side, more time needs to be spent for fine sensing and it increases the probability of interference with other IEEE 802.22 cells in the region.

Contrary to IEEE 802.11af, the decision on the channel to be employed is taken by the base stations themselves. Therefore, such a decision can be made in a collaborative manner meaning that when a cell requires a higher bandwidth to serve some high priority CPEs, so that it needs to aggregate two or three channels, the neighbor cell (which hears the beacons) can decide not to aggregate any channel to give a higher chance to the neighbor cell to find empty (adjacent) channels to be aggregated. Such decisions are made in a differentiated manner, and will be a part of the coexistence mechanisms of the standard. Needless to say that the history-aware recovery scheme can also be employed in the base station, both to refine the sensing lists to be sent to the CPES and to help the base station take the final action.

9.2.3 Cognitive radio in 4G cellular networks

There are some standardization efforts on employing cognitive radios in Wimax under the framework of IEEE 802.16h. The goal behind IEEE 802.16h is to enable coexistence of license-exempt Wimax systems Sherman *et al.* (2008). There are also some proposals to take advantage of cognitive radios in LTE networks Xiao *et al.* (2013); Sun *et al.* (2013). The first suggestion is to use a CR link in TV white spaces as the backhaul link between an LTE base station (eNB) and the LTE core network. This scenario is illustrated in the left part of Figure 9.2. As can be seen, the CR link will be a point-to-point and independent link where the spectrum decisions are made by both endpoints of the link. Therefore, most of the results of this research work are very well applicable to this scenario because they targeted point-to-point and infrastructure networks. When a new channel should be selected, recovery improvement schemes, such as optimal sensing orders and history-awareness, can be used. Considering different cells which are all linked to the core network, we can recognize an

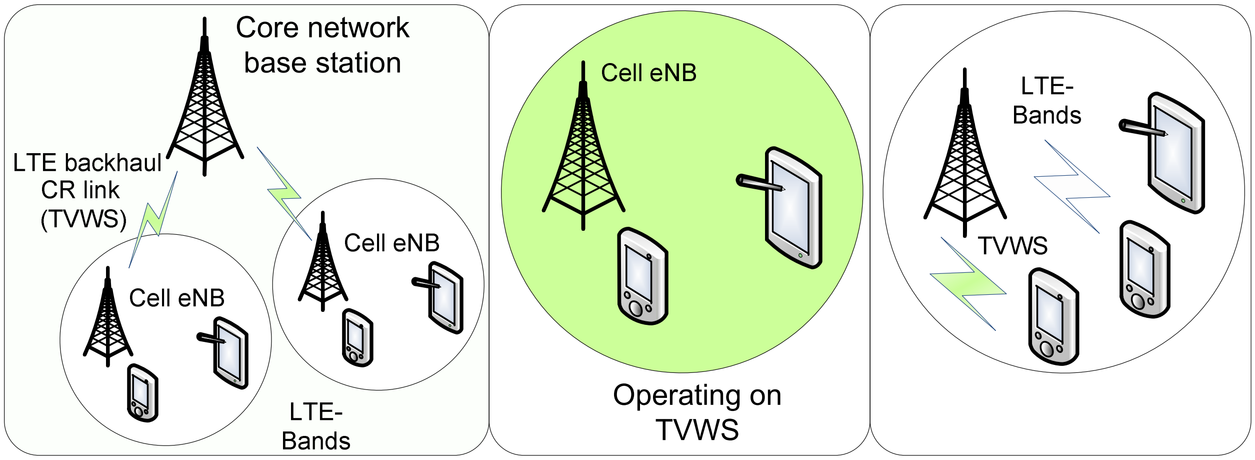

Figure 9.2 Suggested scenarios for the application of cognitive radio in LTE networks. Inspired from the idea in Xiao *et al.* (2013).

infrastructure network composed of the core base station as the access point and cell eNBs as STAs. Each STA operates over a different opportunistic channel, so the network model is the same as the model discussed in Chapter 4. Providing differentiated services (e.g., different priority levels for each cell based on the traffic volume) is straightforward because the core base station has a centralized role and is the main decision-maker.

The next suggestion is to have small LTE cells operating on TV white spaces, as illustrated in the middle part of Figure 9.2. During the normal operation of the network, differentiation can be provided using existing techniques in cellular networks, such as differentiated subcarrier assignment and power adjustment. For the recovery process, it is not realistic to assume that STAs (e.g., mobile phones and tablets) are equipped with sensing capabilities. Therefore, spectrum information will be provided by a central database, similar to IEEE 802.11af, and the same idea of channel selection by an election process may be applicable.

As shown in the right part of Figure 9.2, the last suggestion is to use TV white spaces only as backup channels for the sake of capacity enhancement. In employment of cognitive radios, there will be no major difference with the previous scenario.

CHAPTER 10

CONCLUSION AND FUTURE WORK

With the aim of investigating how reliability and quality of service should be redefined in cognitive radio networks, we first studied the most common sources of failure in wireless networks which resulted in a systematic failure classification procedure in wireless and specifically in cognitive radio networks. It was then explained how cognitive radios can use their inherent capabilities for failure management to implement efficient failure prevention and recovery mechanisms, to combat failures and provide reliable communications and consistent quality of service. It was observed that any cognitive radio feature may be leveraged to improve the reliability while the role of learning and spectrum agility is more significant. The article of Chapter 3 thus opens the way for new designs and evaluation approaches in cognitive radio networks with an aim of improving their reliability.

Considering spectrum agility as the most significant cognitive radio feature and at the same time the current usage of cognitive radio networks and incoming OSA-based standards, the thesis focused on this aspect. This choice can be justified by the fact that what distinguishes a current cognitive radio network from its traditional predecessors is frequent spectrum handovers along with new requirements such as spectrum sensing and spectrum decision. We thus focused on this aspect and modeled the spectrum handover as a failure. Improving the reliability is thus equivalent to increasing mean time to failure, improving the recovery process and shortening the mean time to repair.

We then studied the impact of the recovery time on the reliability and performance of the cognitive radio networks. Failures were classified into hard and soft, then it was investigated how the availability, mean time to failure and mean time to repair for hard and soft failures are affected by the recovery time. We provided closed-form relations for those metrics in a simplified network model which can be used for network design guidelines and in optimization problems. The results of that chapter clearly illustrate how cognitive radios have the potential to provide a reliable wireless network. The results can also be employed at the link level in conjunction with higher level metrics, such as connectivity, to provide a comprehensive reliability model for cognitive radio networks.

It was observed that the time spent for recovery prevents the network from reaching the maximum availability. Therefore, to achieve a high mean time to hard failure and low mean time to repair, the major available option is to increase the number of channels. Thanks to spectrum agility in cognitive radio networks, with a high probability, a user who missed the channel can thus find soon a new channel.

To study the impact of recovery on higher communications layers in presence of a more realistic unsaturated traffic, a queueing approach was chosen. Recovery periods were considered as a service interruption, so that we proposed a general M/G/1 queueing model with interruption, which can be employed in different scenarios and can address a general class of network models. Different reliability and quality of service parameters have been found from this queueing model to investigate how channel parameters, such as availability and unavailability periods, and the recovery algorithm parameters, such as the recovery duration, affect the average delay. Derived closed-form equations enable us to evaluate any new proposed recovery model by deriving first the recovery periods' distribution and then applying the queueing models.

To support traffic differentiation and to investigate the impact of recovery periods on the joint performance of multiple classes of traffic, we suggested a priority queueing approach. We extended the approaches of the general queueing model and discussed four different priority queueing disciplines ranging from a pure preemptive scheme to a pure non-preemptive scheme. New disciplines increase the flexibility and decision resolution and enable the CR node to have a more accurate control of the interaction between different classes of traffic. The models are solved, so it can be discussed how the reliability and quality of service parameters, such as delay and jitter, for a specific class of traffic are affected not only by the channel parameters, but also by the characteristics of other traffic classes.

The M/G/1 queueing model with interruptions was a foundation for performance analysis and an answer to the need of having closed-form analytical relations. We then extended the queueing model to more practical scenarios, first with heterogeneous channels and second with multiple users and a random medium access model. In the first part, the queue occupancy was modeled as a multi-row Markov chain where each row represents one of the possible service rates. In addition to numerically solving the Markov chain, two analytical approximations are provided. We obtained a queueing model to analyze cognitive radio performance, in an OSA network with heterogeneous operating channels, so the results can be used in optimization models and cognitive radio network design.

In the next chapter, a delay analysis was performed for a multi-user scenario with discrete-time distributions to have more insights on the impact of a baseline random medium access protocol and design of control channel on the performance of the cognitive radio networks. We did a comprehensive investigation on the interaction of recovery algorithms and the medium access protocol and provided a comprehensive analytical delay analysis for a multichannel cognitive radio network with both buffering and switching recovery policies where the impact of different network parameters on the performance is considered. The results can thus be used in network optimization and design problems. It was observed that with an Aloha-type medium access, the average delay is lower when instead of a switching recovery policy (i.e., switching to a new channel in case of appearance of primary users), a waiting and buffering recovery policy is employed (i.e., the CR user waits for the primary user to vacate the channel) because in an Aloha-type medium access with a common control channel, the bottleneck is access to control channel and a switching policy necessitates more frequent control channel accesses.

While the research work was mostly focused on analyzing the impact of spectrum handover recovery process, in the last chapter, it was discussed how learning and history-awareness in cognitive radio networks can be used to improve the recovery process. A greedy and history-aware spectrum handover scheme was suggested to improve the time spent for spectrum handover. To perform spectrum handover in an optimal manner, at the beginning of each timeslot and based on the state of the current channel, the scheme computes the optimal number of channels to be sensed where and this number is dynamically updated. On the other side, intrinsic features of learning and history-awareness of CRs were used to create an optimal list of channels to be sensed based on the channels' background and historical information. Simulation results showed that the history-aware sensing order improves the restoration mechanism by providing a shorter restoration time or a restored channel with a higher quality. The proposed history-aware greedy scheme is general and applicable to many communication systems. It was discussed, for instance, how it can be integrated into LTE networks considering MIMO and in-band channel aggregation features of LTE-advanced.

10.1 Limitations and future work

The objective of this research work was to investigate how reliability and quality of service can be evaluated and increased in a cognitive radio network based on opportunistic spectrum access. Due to the recentness of the research in this area, there is a lack of a rich literature on reliability and quality of service related to cognitive networks. Thus, part of the future work should be applying and extending the suggested solutions to differentiated scenarios.

In the performance evaluation sections, both for saturated and unsaturated traffic, we could not consider all scenarios and possibilities. For instance, priority queueing models were analyzed for a service-resume model and continuous distributions. Extension to service-repeat models and discrete-time (time-slotted) distributions are parts of the future work. Further investigation is also required to find other performance metrics such as packet loss (finite buffer), MTTF and MTTR for hard and soft failures from the queueing models.

When it comes to multi-user scenarios, cognitive radios are still in their prototyping and evaluation period. On the one hand, there is no approved and implemented standard or protocol for CR networking and medium access, and on the other hand, there are many suggestions and proposals in the literature. Instead of focusing on a specific proposal or model, we decided to propose and discuss models and tools which are more general, so that they can be applicable, with slight changes, to any medium access protocol. In the future, with the advent of more approved protocols and standards, the results of this research work can be revisited to employ them in specific realistic and practical scenarios.

The MAC model discussed in Chapter 7 was a baseline to investigate two possible spectrum handover policies. Aloha has thus been used for simplicity since the focus of the work was the spectrum handover and not the medium access and contentions. However, for future work, the work can be extended to more realistic models, for instance, based on Carrier Sense Multiple Access (CSMA). The collision between secondary users can also be addressed by taking into account a state-dependent transmission success meaning that the transmission success will be variable, depending on the number of transmitting users, with the price of increased complexity. By increasing the complexity of state representation (e.g., more state variables), it will also be possible to address more realistic non-memoryless distributions for the channel model and packet length. It was also assumed in the same chapter that the list of available channels is known. The multi-channel hidden problem which addresses the cases where this knowledge is not perfectly available remains also as a future work.

Recalling from Chapter 3, we saw that cognitive radios are powerful and well-equipped radios which can participate in the improvement of reliability or quality of service in various ways. In this research work, we mostly focused on the capabilities of spectrum-awareness and frequency switching in CRNs and establish our models based on these capabilities. In other words, the objective was always improving the recovery process with optimizing the sensing orders. However, other features of cognitive radios such as learning and location-awareness can also play important roles in the reliability of cognitive wireless networks and should be explored in future work.

10.1.1 Multihop cognitive radio networks

From the topology point of view, most of the discussions in this research work have been applicable to a single-hop (single link) cognitive radio network ; i.e., a point-to-point communication between two nodes or an infrastructure network. The reason was twofold. First, even for a multi-hop cognitive radio network, the starting point is the single link. Therefore, what has been done in this research work is a building block for further research on multi-hop cognitive radio networks. For instance, suggested history-aware model for channel selection will be employed in all nodes of a multi-hop network. Second, we can see that all standardization efforts in the area of cognitive radio networks and dynamic spectrum access have also been for infrastructure networks such that both standards under development based on cognitive radio networks, IEEE 802.22 Cordeiro *et al.* (2006) and IEEE 802.11af Flores *et al.* (2013), are infrastructure networks, so the results of this research work can easily be extended and applied to practical scenarios. Naturally, reliability improvement in multi-hop cognitive radio networks, where features such as location-awareness and decision-making on joint spectrum selection and routing Sengupta et Subbalakshmi (2013) may play a determining role, remains as a future work. The proposed research in this area can be categorized as :

- Applying the queuing model to MHCRNs ;
- Minimizing channel searching time similar to our discussion in Chapter 8 ;
- Differentiated routing and channel assignment, assuming perfect spectrum knowledge.

Network of queues

Given the queueing results obtained in this thesis, the possibility of using the existing methods, such as the Kleinrock Independence Approximation and extensions of Jackson's Theorem Bertsekas et Gallager (1992), for a network of queues should be investigated. The motivation is the fact that in MHCRNs, due to the variations of the spectrum, the notion of opportunistic or probabilistic routing is very well applicable. That is, instead of using a pre-defined path, a node may route the packets randomly to any of the eligible neighbors. For this end, we first consider a single middle node in the network as follows. As illustrated in Figure 10.1, assume a CR node which is the next hop of M_p CR nodes (predecessors) and relays the received traffic to M_s CR nodes (successors). The channel which is being used to communicate to each predecessor or successor could be different thus a maximum of $M_s + M_p$ channels are being used. The priority of each flow is different which implies that we may

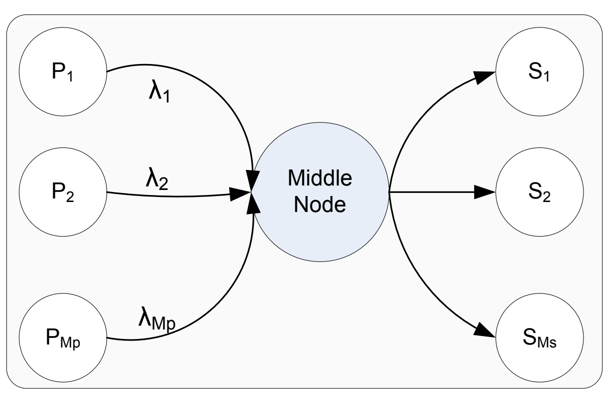

Figure 10.1 A relay node in an MHCRN.

have M_p different priority classes of flows. A differentiated scheduling scheme needs to be discussed and then the queuing relations can be derived.

Differentiated channel selection

In an infrastructure network, it is assumed that all nodes send their sensing results to the access point (AP) and the AP creates the sensing order in a centralized manner. In a multi-hop (and usually ad-hoc) CRN, none of these assumptions is necessarily valid ; therefore, the objective is first to propose a method for distributed learning then using this method inside a framework for distributed channel selection. The preliminary idea is as follows.

In a recovery period, there are some break periods where the users broadcast their sensing results over the control channel. Their neighbors hear this message and analyze that to update their spectrum knowledge for future use which is the base of distributed learning. Moreover, if they have an available common channel with either of the endpoints of a broken link, they inform them. These proposed common channels can be used for local path recovery instead of a link recovery which implies that the broken link will be recovered through another middle hop. The endpoints of a broken link can decide to accept this offer and accept the relay path, or continue their channel sensing with the hope of finding a common channel for their local link recovery. There will be some trade-offs to discuss ; for instance, the decision between continuing the local recovery or accepting the path recovery offer, or decision of the broadcast range of the sensing results. If the nodes broadcast their sensing results to more neighbors, their chance of path recovery increases. However, the overhead of signaling over the control channel also increases.

For the sake of differentiation, recovery priority should be defined based on the priority of the broken flows. The idea of preemption can be applied using overhearing. When two nodes confirm using a new channel, they inform the neighbors by broadcasting. Another node with a higher-priority broken flow hears this message and may claim the channel.

Differentiated routing and channel assignment

For MHCRNs with a perfect spectrum knowledge, the recovery is equivalent to a routing and channel assignment problem where we are looking for a new route and channel for the broken flows with different priorities. The best recovery decision (link recovery versus path recovery, local versus global and preemption versus no-preemption) considering the constraints such as the flow size, flow priority, interference, capacity of the links (quality of the channels) and timing costs of channel switching and preemption should be made. Several variations of this problem can be considered and discussed. For instance :

- Centralized or distributed ;
- Possibility of preemption (changing the existing flows' route or channels) ;
- Level of recovery : link, local path or global path recovery
- Possibility of bifurcation.

This problem has been discussed in the literature for wireline networks and also for cognitive radio networks without differentiation. Some ideas for differentiation are as follows.

- Differentiated route metrics : We have usually different requirements for different types of traffic. Inside a routing protocol, this is equivalent to different route metrics. Three main metrics which should be considered are delay, throughput and stability (less switching) ;
- Preemption : A flow with a higher priority can preempt the resource which is being used by a flow with a lower priority ;
- Recovery priority : When a failure occurs and some affected flows should be re-routed, the recovery procedure is started first for a flow with the highest priority (more applicable to distributed schemes).

10.1.2 Wideband (Underlay) cognitive radio networks

Finally, we mainly considered a narrowband communication system, known as *interweave* in the literature of cognitive radios Goldsmith *et al.* (2009), where each channel is being used by one user, either the cognitive radio user or the primary user. This research can be extended to wideband communication schemes where cognitive radio users and primary users coexist and the CR power adjustment to maintain an acceptable interference level at the primary user will be a key solution. When multiple users, either several cognitive users or the cognitive and primary users, can work on the same channel, the definition of reliability should be modified, as discussed in Chapter 3, because in that case the cognitive radio user does not necessarily vacate the channel. It may decrease its power and remain on the same channel. It is worth noting that the channel model proposed in Chapter 8 remains valid because power adjustment is considered to be a change in the channel state. In other words, the transition probabilities between channel states will not only be a function of the fading process, but also of the variation of transmission power.

REFERENCES

AKAN, O., KARLI, O. et ERGUL, O. (2009). Cognitive radio sensor networks. *IEEE Network*, 23, 34–40.

AKYILDIZ, I., WANG, X. et WANG, W. (2005). Wireless mesh networks : a survey. *Computer Networks*, 47, 445–487.

AKYILDIZ, I. F., LEE, W. et CHOWDHURY, K. R. (2009). CRAHNs : Cognitive radio ad hoc networks. *Ad Hoc Networks*, 7, 810 – 836.

AKYILDIZ, I. F., LEE, W.-Y., VURAN, M. C. et MOHANTY, S. (2006). Next generation/dynamic spectrum access/cognitive radio wireless networks : A survey. *Computer Networks*, 50, 2127 – 2159.

AL-KUWAITI, M., KYRIAKOPOULOS, N. et HUSSEIN, S. (2009). A comparative analysis of network dependability, fault-tolerance, reliability, security, and survivability. *IEEE Communications Surveys & Tutorials*, 11, 106 –124.

ASTELY, D., DAHLMAN, E., FURUSKAR, A., JADING, Y., LINDSTROM, M. et PARKVALL, S. (2009). LTE : the evolution of mobile broadband. *IEEE Communications Magazine*, 47, 44 –51.

AVI-ITZHAK, B. (1963). Preemptive repeat priority queues as a special case of the multipurpose server problem –II. *Operations Research*, 11, 610–619.

AZARFAR, A., FRIGON, J.-F. et SANSÒ, B. (2010). A performance comparison of cognitive versus traditional radio networks. *Proc. of the IEEE Globecom 2010 Workshop on Broadband Wireless Access*. 778–782.

AZARFAR, A., FRIGON, J.-F. et SANSÒ, B. (2011a). A greedy channel restoration scheme for opportunistic spectrum access. *2011 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*. 476–481.

AZARFAR, A., FRIGON, J.-F. et SANSÒ, B. (2011b). History-aware channel search schemes in cognitive radio networks. *IEEE International Symposium in Personal, Indoor and Mobile Radio Communications, PIMRC 2011*. 379–383.

AZARFAR, A., FRIGON, J.-F. et SANSÒ, B. (2012a). Analysis of cognitive radio networks based on a queueing model with server interruptions. *Proc. of IEEE International Conference on Communications, ICC 2012*. 1703–1708.

AZARFAR, A., FRIGON, J.-F. et SANSÒ, B. (2012b). Dynamic selection of priority queueing discipline in cognitive radio networks. *Proc. of IEEE 76th Vehicular Technology Conference (VTC2012-Fall)*. 1–5.

AZARFAR, A., FRIGON, J.-F. et SANSÒ, B. (2012c). Improving the reliability of wireless networks using cognitive radios. *IEEE Communications Surveys and Tutorials*, 14, 338–354.

AZARFAR, A., FRIGON, J.-F. et SANSÒ, B. (2013a). Optimal sensing order in cognitive radio networks considering channel stability and traffic differentiation. *Proc. of IEEE Globecom 2013*. 1125–1130.

AZARFAR, A., FRIGON, J.-F. et SANSÒ, B. (2013b). Reliability analysis of a channel restoration mechanism for opportunistic spectrum access. *Computer Communications*, 36, 947 – 961.

AZARFAR, A., FRIGON, J.-F. et SANSÒ, B. (2014a). Delay analysis of multichannel opportunistic spectrum access mac protocols. Submitted to IEEE Transactions on Mobile Computing. [Available Online] at http ://arxiv.org/pdf/1403.7192v1.pdf.

AZARFAR, A., FRIGON, J.-F. et SANSÒ, B. (2014b). Hierarchical channel recovery for heterogeneous cognitive radio networks. *Proc. of IEEE DYSPAN 2014*. 145–156.

AZARFAR, A., FRIGON, J.-F. et SANSÒ, B. (2014c). A history-aware greedy channel restoration scheme for cognitive radio-based {LTE} networks. *Physical Communication*, 10, 116 – 126.

AZARFAR, A., FRIGON, J.-F. et SANSÒ, B. (2014d). Priority queueing models for cognitive radio networks with traffic differentiation. Submitted to EURASIP Journal on Wireless Communications and Networking.

AZARFAR, A., FRIGON, J.-F. et SANSÒ, B. (2014e). Queueing model for heterogeneous opportunistic spectrum access. Submitted to IET Communications.

BALL, M., COLBOURN, C. et PROVAN, J. (1995). *Network Models*, Elsevier, chapitre 11.—Network Reliability. Handbooks in Operations Research and Management Science.

BANY SALAMEH, H. et KRUNZ, M. (2009). Channel access protocols for multihop opportunistic networks : challenges and recent developments. *IEEE Network*, 23, 14 –19.

BARLOW, R. E., PROSCHAN, F. et HUNTER, L. C. (1965). *Mathematical Theory of Reliability*. John Wiley & Sons, Inc.

BERTHOLD, U., F., F., VAN DER SCHAAR, M. et JONDRAL, F. (2008). Detection of spectral resources in cognitive radios using reinforcement learning. *Proc. of 3rd IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks, 2008. DySPAN 2008.*

BERTSEKAS, D. et GALLAGER, R. (1992). *Data Networks*. Prentice-Hall, seconde édition.

BERTSEKAS, D. P. (2005). *Dynamic Programming and Optimal Control*. Athena Scientific. Vol. I, 3rd Edition.

BIANCHI, G. (2000). Performance analysis of the IEEE 802.11 distributed coordination function. *IEEE Journal on Selected Areas in Communications*, 18, 535–547.

BICEN, A. O., GUNGOR, V. C. et AKAN, O. B. (2012). Delay-sensitive and multimedia communication in cognitive radio sensor networks. *Ad Hoc Networks*, 10, 816 – 830. Special Issue on Cognitive Radio Ad Hoc Networks.

CABRIC, D., MISHRA, S. et BRODERSEN, R. (2004). Implementation issues in spectrum sensing for cognitive radios. *Conference Record of the Thirty-Eighth Asilomar Conference on Signals, Systems and Computers, 2004.* vol. 1, 772 – 776 Vol.1.

CHANG, N. B. et LIU, M. (2007). Optimal channel probing and transmission scheduling for opportunistic spectrum access. *Proc. of MobiCom '07.* 27–38.

CHEN, D., YIN, S., ZHANG, Q., LIU, M. et LI, S. (2009). Mining spectrum usage data : a large-scale spectrum measurement study. *Proc. of MobiCom '09.* 13–24.

CHENG, G., LIU, W., LI, Y. et CHENG, W. (2007). Joint on-demand routing and spectrum assignment in cognitive radio networks. *IEEE International Conference on Communications, 2007(ICC '07).* 6499 –6503.

CHIA-CHUN HSU, A., WEIT, D. et KUO, C.-C. (2007). A cognitive MAC protocol using statistical channel allocation for wireless ad-hoc networks. *Proc. of IEEE Wireless Communications and Networking Conference, 2007 (WCNC 2007).* 105–110.

CHOI, S., DEL PRADO, J., N, S. S. et MANGOLD, S. (2003). IEEE 802.11 e contention-based channel access (EDCF) performance evaluation. *Proc. of IEEE International Conference on Communications, 2003. ICC '03.* vol. 2, 1151 –1156 vol.2.

CHOLDA, P., MYKKELTVEIT, A., HELVIK, B., WITTNER, O. et JAJSZCZYK, A. (2007). A survey of resilience differentiation frameworks in communication networks. *IEEE Communications Surveys & Tutorials*, 9, 32 –55.

CHOLDA, P., TAPOLCAI, J., CINKLER, T., WAJDA, K. et JAJSZCZYK, A. (2009). Quality of resilience as a network reliability characterization tool. *IEEE Network*, 23, 11 –19.

CHOWDHURY, K. et AKYILDIZ, I. (2008). Cognitive wireless mesh networks with dynamic spectrum access. *IEEE Journal on Selected Areas in Communications*, 26, 168–181.

CISCO (2013). Cisco visual networking index : Global mobile data traffic forecast update, 2012—2017. Cisco White Paper.

CLANCY, C., HECKER, J., STUNTEBECK, E. et O'SHEA, T. (2007). Applications of machine learning to cognitive radio networks. *IEEE Wireless Communications,*, 14, 47 –52.

CORDEIRO, C. et CHALLAPALI, K. (2007). C-MAC : A cognitive MAC protocol for multi-channel wireless networks. *2nd IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks, 2007. DySPAN 2007.* 147–157.

CORDEIRO, C., CHALLAPALI, K., BIRRU, D. et SAI SHANKAR, N. (2006). IEEE 802.22 : An introduction to the first wireless standard based on cognitive radios. *Journal of Communications*, 1, 38–47.

CORMIO, C. et CHOWDHURY, K. R. (2009). A survey on MAC protocols for cognitive radio networks. *Ad Hoc Networks*, 7, 1315–1329.

COX, D. R. (1962). *Renewal Theory.* Methuen ; Wiley London, New York.

D. P. GAVER, J. (1962). A waiting line with interrupted service, including priorities. *Journal of the Royal Statistical Society. Series B (Methodological)*, 24, 73–90.

DEMEESTER, P., GRYSEELS, M., AUTENRIETH, A., BRIANZA, C., CASTAGNA, L., SIGNORELLI, G., CLEMENTE, R., RAVERA, M., JAJSZCZYK, A., JANUKOWICZ, D., DOORSELAERE, K. et HARADA, Y. (1999). Resilience in multilayer networks. *IEEE Communications Magazine*, 37, 70–76.

DILLINGER, M., K.MADANI et ALONISTIOTI, N. (2003). *Software Defined Radio : Architectures, Systems and Functions.* John Wiley & Sons, Chichester, UK.

DING, L., MELODIA, T., BATALAMA, S., MATYJAS, J. et MEDLEY, M. (2010). Cross-layer routing and dynamic spectrum allocation in cognitive radio ad hoc networks. *IEEE Transactions on Vehicular Technology,*, 59, 1969 –1979.

DOYLE, L., DASILVA, L., OZGUL, B. et FORDE, T. (2011). A letter from the editors. Elsevier Physical Communications, Call for Papers for Special Issue on Cognitive Radio for LTE Advanced and Beyond.

ET AL., P. P. (2005). RFC4090-fast reroute extensions to RSVP-TE for LSP tunnels.

FAN, R. et JIANG, H. (2009). Channel sensing-order setting in cognitive radio networks : A two-user case. *IEEE Transactions on Vehicular Technology*, 58, 4997 –5008.

FEDERGRUEN, A. et GREEN, L. (1986). Queueing systems with service interruptions. *Operations Research*, 34, 752–768.

FETTE, B. A. et FETTE, B. (2006). *Cognitive Radio Technology (Communications Engineering).* Newnes.

FIEMS, D., MAERTENS, T. et BRUNEEL, H. (2008). Queueing systems with different types of server interruptions. *European Journal of Operational Research*, 188, 838–845.

FLORES, A., GUERRA, R., KNIGHTLY, E., ECCLESINE, P. et PANDEY, S. (2013). Ieee 802.11af : a standard forTV white space spectrum sharing. *IEEE Communications Magazine*, 51, 92–100.

GABRAN, W., LIU, C., PAWELCZAK, P. et CABRIC, D. (2013). Primary user traffic estimation for dynamic spectrum access. *IEEE Journal on Selected Areas in Communications*, 31, 544–558.

GEIRHOFER, S., T., L. et SADLER, B. (2008). Cognitive medium access : Constraining interference based on experimental models. *IEEE Journal on Selected Areas in Communications*, 26, 95–105.

GHASEMI, A. et SOUSA, E. (2005). Collaborative spectrum sensing for opportunistic access in fading environments. *First IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks, 2005 (DySPAN 2005)*. 131 –136.

GHASEMI, A. et SOUSA, E. (2008). Spectrum sensing in cognitive radio networks : requirements, challenges and design trade-offs. *IEEE Communications Magazine*, 46, 32 –39.

GOLDSMITH, A., JAFAR, S., MARIC, I. et SRINIVASA, S. (2009). Breaking spectrum gridlock with cognitive radios : An information theoretic perspective. *Proceedings of the IEEE*, 97, 894–914.

GRANELLI, F., PAWELCZAK, P., PRASAD, R., SUBBALAKSHMI, K., CHANDRAMOULI, R., HOFFMEYER, J. et BERGER, H. (2010). Standardization and research in cognitive and dynamic spectrum access networks : IEEE SCC41 efforts and other activities. *IEEE Communications Magazine*, 48, 71 –79.

GROVER, W. (2003). *Mesh-based Survivable Transport Networks : Options and Strategies for Optical, MPLS, SONET and ATM Networking*. Prentice Hall PTR.

HAIDER, A. et HARRIS, R. (2007). Recovery techniques in next generation networks. *IEEE Communications Surveys & Tutorials*, 9, 2 –17.

HAN, Z., JI, Z. et LIU, K. (2005). Fair multiuser channel allocation for OFDMA networks using nash bargaining solutions and coalitions. *IEEE Transactions on Communications*, 53, 1366 – 1376.

HAYKIN, S. (2005). Cognitive radio : Brain-empowered wireless communications. *IEEE Journal on Selected Areas in Communications,*, 23, 201–220.

HUANG, S., LIU, X. et DING, Z. (2008). Opportunistic spectrum access in cognitive radio networks. *Proc. of The 27th Conference on Computer Communications. IEEE INFOCOM 2008.* 1427–1435.

IEEE (2004). IEEE standard for mac bridges. Standard IEEE 802.1D-2004.

IEEE (2008). IEEE standard for link aggregation. Standard IEEE 802.1AX-2008.

JIA, J., ZHANG, Q. et SHEN, X. (2008). HC-MAC : A hardware-constrained cognitive MAC for efficient spectrum management. *IEEE Journal on Selected Areas in Communications*, 26, 106 –117.

JIANG, H., LAI, L., FAN, R. et POOR, H. (2009). Optimal selection of channel sensing order in cognitive radio. *IEEE Transactions on Wireless Communications*, 8, 297 –307.

JONDRAL, F. (2007). Cognitive radio : A communications engineering view. *IEEE Wireless Communications*, 14, 28 –33.

JONDRAL, F. K. (2005). Software-defined radio : basics and evolution to cognitive radio. *EURASIP J. Wirel. Commun. Netw.*, 2005, 275–283.

KANT, L. et CHEN, W. (2005). Service survivability in wireless networks via multi-layer self-healing. *Proc. of IEEE Wireless Communications and Networking Conference 2005 (WCNC'05)*. vol. 4, 2446–2452.

KHALIFE, H., MALOUCH, N. et FDIDA, S. (2009). Multihop cognitive radio networks : to route or not to route. *IEEE Network*, 23, 20 –25.

KIM, H. et SHIN, K. (2008a). Efficient discovery of spectrum opportunities with mac-layer sensing in cognitive radio networks. *IEEE Transactions on Mobile Computing*, 7, 533 –545.

KIM, H. et SHIN, K. (2008b). Fast discovery of spectrum opportunities in cognitive radio networks. *Proc. of 3rd IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks, 2008. DySPAN 2008.* 1–12.

KIM, K. (2012). T-preemptive priority queue and its application to the analysis of an opportunistic spectrum access in cognitive radio networks. *Comput. Oper. Res.*, 39, 1394–1401.

KLEINROCK, L. (1975). *Queuing Systems V.1.* Wiley.

LAI, L., GAMAL, H. E., JIANG, H. et POOR, H. V. (2011). Cognitive medium access : Exploration, exploitation, and competition. *IEEE Transactions on Mobile Computing*, 10, 239–253.

LAOURINE, A., CHEN, S. et TONG, L. (2010). Queuing analysis in multichannel cognitive spectrum access : A large deviation approach. *Proc. of IEEE INFOCOM 2010.* 1 –9.

LI, H. et HAN, Z. (2011). Socially optimal queuing control in cognitive radio networks subject to service interruptions : To queue or not to queue ? *IEEE Transactions on Wireless Communications*, 10, 1656 –1666.

LI, H. et QIAN, L. (2010). Enhancing the reliability of cognitive radio networks via channel assignment : risk analysis and redundancy allocation. *Proc. of 44th Annual Conference on Information Sciences and Systems (CISS), 2010*. 1–6.

LI, K.-F., LAU, W.-C. et YUE, O.-C. (2008). Link restoration in cognitive radio networks. *Proc. of IEEE International Conference on Communications 2008 (ICC'08)*. 371–376.

LIANG, Y., ZENG, Y., PEH, E. et HOANG, A. T. (2008). Sensing-throughput tradeoff for cognitive radio networks. *IEEE Transactions on Wireless Communications*, 7, 1326–1337.

LIU, H., ZHANG, B., MOUFTAH, H., SHEN, X. et MA, J. (2009). Opportunistic routing for wireless ad hoc and sensor networks : Present and future directions. *IEEE Communications Magazine*, 47, 103 –109.

LIU, K., ZHAO, Q. et CHEN, Y. (2008). Distributed sensing and access in cognitive radio networks. *IEEE ISSSTA '08*. 23–27.

LUO, L. et ROY, S. (2007). Analysis of search schemes in cognitive radio. *2nd IEEE Workshop on Networking Technologies for Software Define Radio Networks*. 17 –24.

MA, M. et TSANG, D. H. K. (2008). Impact of channel heterogeneity on spectrum sharing in cognitive radio networks. *Proc. of IEEE International Conference on Communications, 2008. ICC '08*. 2377–2382.

MACKENZIE, A., REED, J., ATHANAS, P., BOSTIAN, C., BUEHRER, R., DASILVA, L., ELLINGSON, S., HOU, Y., HSIAO, M., PARK, J.-M., PATTERSON, C., RAMAN, S. et DA SILVA, C. (2009). Cognitive radio and networking research at virginia tech. *Proceedings of the IEEE*, 97, 660 –688.

MAHMOUD, H., YUCEK, T. et ARSLAN, H. (2009). OFDM for cognitive radio : merits and challenges. *IEEE Wireless Communications*, 16, 6 –15.

MAHMOUD, Q. (2007). *Cognitive Networks : Towards Self-Aware Networks*. Wiley-Interscience.

MAUVE, M., WIDMER, A. et HARTENSTEIN, H. (2001). A survey on position-based routing in mobile ad hoc networks. *IEEE Network*, 15, 30 –39.

MITOLA, J. (2009). Cognitive radio architecture evolution. *Proceedings of the IEEE*, 97, 626 –641.

MITOLA III, J. (2000). *Cognitive Radio : An Integrated Agent Architecture for Software Defined Radio*. Thèse de doctorat, Royal Institute of Technology, Stockholm, Sweeden.

MITOLA III, J. (2001). Cognitive radio for flexible mobile multimedia communications. *Mobile Networks and Applications*, 6, 435–441. 10.1023/A :1011426600077.

MO, J., SO, H.-S. et WALRAND, J. (2008). Comparison of multichannel MAC protocols. *IEEE Transactions on Mobile Computing*, 7, 50 –65.

NISAN, N., ROUGHGARDEN, T., TARDOS, E. et VAZIRANI, V. (2007). *Algorithmic Game Theory*. Cambridge University Press, New York, NY, USA.

NIYATO, D. et HOSSAIN, E. (2009). Cognitive radio for next-generation wireless networks : an approach to opportunistic channel selection in IEEE 802.11-based wireless mesh. *IEEE Wireless Communications*, 16, 46–54.

NOJO, S. et WATANABE, H. (1993). Incorporating reliability specifications in the design of telecommunication networks. *IEEE Communications Magazine*, 31, 40 –43.

PAL, R. (2007). On the reliability of multi-hop dynamic spectrum access networks supporting QoS driven applications. *Proc. of IEEE International Conference on Communications, 2007. ICC '07*. 5294–5299.

PARK, J., PAWECZAK, P. et CABRIC, D. (2011). Performance of joint spectrum sensing and MAC algorithms for multichannel opportunistic spectrum access ad hoc networks. *IEEE Transactions on Mobile Computing*, 10, 1011–1027.

PAWELCZAK, P., POLLIN, S., SO, H., BAHAI, A. R. S., V. PRASAD, R. et HEKMAT, R. (2009). Performance analysis of multichannel medium access control algorithms for opportunistic spectrum access. *IEEE Transactions on Vehicular Technology*, 58, 3014–3031.

PAWELCZAK, P., POLLIN, S., SO, H.-S., BAHAI, A., PRASAD, R. et HEKMAT, R. (2008). Quality of service assessment of opportunistic spectrum access : a medium access control approach. *IEEE Wireless Communications*, 15, 20 –29.

PELUSI, L., PASSARELLA, A. et CONTI, M. (2006). Opportunistic networking : data forwarding in disconnected mobile ad hoc networks. *IEEE Communications Magazine*, 44, 134 –141.

RAPPAPORT, T. (2001). *Wireless Communications : Principles and Practice*. Prentice Hall PTR, Upper Saddle River, NJ, USA.

RASHID, M., HOSSAIN, J., HOSSAIN, E. et BHARGAVA, V. (2007). Opportunistic spectrum access in cognitive radio networks : A queueing analytic model and admission controller design. *Proc. of IEEE GLOBECOM '07*. 4647–4652.

RASHID, M., HOSSAIN, M., HOSSAIN, E. et BHARGAVA, V. (2009). Opportunistic spectrum scheduling for multiuser cognitive radio : a queueing analysis. *IEEE Transactions on Wireless Communications*, 8, 5259–5269.

ROSS, S. M. (2006). *Introduction to Probability Models, Ninth Edition*. Academic Press, Inc., Orlando, FL, USA.

SAATSAKIS, A., TSAGKARIS, K., VON HUGO, D., SIEBERT, M., ROSENBERGER, M. et DEMESTICHAS, P. (2008). Cognitive radio resource management for improving the efficiency of LTE network segments in the wireless B3G world. *Proc. of 3rd IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks, DySPAN 2008.*

SACHS, J., MARIC, I. et GOLDSMITH, A. (2010). Cognitive cellular systems within the tv spectrum. *2010 IEEE Symposium on New Frontiers in Dynamic Spectrum*. 1 –12.

SANSÒ, B., AWAD, C. et GIRARD, A. (2006). Can DiffServ guarantee IP QoS under failures ? *IEEE Network*, 20, 32–40.

SANSÒ, B. et SOUMIS, F. (1991). Communications and transportation networks reliability using routing models. *IEEE Transactions on Reliability*, 29–38.

SEN, A. (1986). Chapter 22 : Social choice theory. Elsevier, vol. 3 de *Handbook of Mathematical Economics*. 1073 – 1181.

SENGUPTA, S. et SUBBALAKSHMI, K. (2013). Open research issues in multi-hop cognitive radio networks. *IEEE Communications Magazine*, 51, 168–176.

SHANKAR, N. S. (2007). Efficiency and coexistence strategies for cognitive radio. H. Arslan, éditeur, *Cognitive Radio, Software Defined Radio, and Adaptive Wireless Systems*, Springer Netherlands, Boston/Dordrecht/London, chapitre 7. 189–234.

SHERMAN, M., MODY, A., MARTINEZ, R., RODRIGUEZ, C. et REDDY, R. (2008). IEEE standards supporting cognitive radio and networks, dynamic spectrum access, and coexistence. *IEEE Communications Magazine*, 46, 72–79.

SHIANG, H.-P. et VAN DER SCHAAR, M. (2008). Queuing-based dynamic channel selection for heterogeneous multimedia applications over cognitive radio networks. *IEEE Trans. on Multimedia*, 10, 896–909.

SHOOMAN, M. L. (2002). *Reliability of Computer Systems and Networks : Fault Tolerance, Analysis, and Design*. John Wiley & Sons, Inc., New York, USA.

SIEWIOREK, D. P. et SWARZ, R. S. (1998). *Reliable computer systems (3rd ed.) : design and evaluation*. A. K. Peters, Ltd., Natick, MA, USA.

SNOW, A. P., VARSHNEY, U. et MALLOY, A. D. (2000). Reliability and survivability of wireless and mobile networks. *Computer*, 33, 49–55.

STEVENSON, C., CHOUINARD, G., LEI, Z., HU, W., SHELLHAMMER, S. et CALDWELL, W. (2009). IEEE 802.22 : The first cognitive radio wireless regional area network standard. *IEEE Communications Magazine*, 47, 130–138.

STOJMENOVIC, I. (2002). Position-based routing in ad hoc networks. *IEEE Communications Magazine*, 40, 128 –134.

SU, H. et ZHANG, X. (2008). Cross-layer based opportunistic MAC protocols for QoS provisionings over cognitive radio wireless networks. *IEEE Journal on Selected Areas in Communications*, 26, 118–129.

SUN, S., JU, Y. et YAMAO, Y. (2013). Overlay cognitive radio OFDM system for 4G cellular networks. *IEEE Wireless Communications*, 20, 68–73.

TAKAGI, H. (1991). *Queueing analysis : a foundation of performance evaluation, Volume 1*. North-Holland. North-Holland.

TAKAGI, H. (1993). *Queueing Analysis : Discrete-time systems*. Queueing Analysis : A Foundation of Performance Evaluation. North-Holland.

TIPPER, D., DAHLBERG, T., SHIN, H. et CHARNSRIPINYO, C. (2002). Providing fault tolerance in wireless access networks. *IEEE Communications Magazine*, 40, 58–64.

TSE, D. et VISWANATH, P. (2005). *Fundamentals of Wireless Communications*. Cambridge University Press.

ULVERSOY, T. (2010). Software defined radio : Challenges and opportunities. *IEEE Communications Surveys & Tutorials*, 12, 531–550.

VASSEUR, J.-P., PICKAVET, M. et DEMEESTER, P. (2004). *Network Recovery : Protection and Restoration of Optical, SONET-SDH, IP, and MPLS*. Morgan Kauffman.

VUČEVIĆ, N., PÉREZ-ROMERO, J., SALLENT, O. et AGUSTÍ, R. (2011). Reinforcement learning for joint radio resource management in LTE-UMTS scenarios. *Computer Networks*, 55, 1487 – 1497. Recent Advances in Network Convergence.

W. LAI, D. M. (2002). RFC3386-network hierarchy and multilayer survivability.

WANG, L., WANG, C. et FENG, K. (2011). A queueing-theoretical framework for QoS-enhanced spectrum management in cognitive radio networks. *IEEE Wireless Communications*, 18, 18 –26.

WANG, S., ZHANG, J. et TONG, L. (2010). Delay analysis for cognitive radio networks with random access : A fluid queue view. *Proc. of IEEE INFOCOM 2010*. 1–9.

WANG, S., ZHANG, J. et TONG, L. (2012a). A characterization of delay performance of cognitive medium access. *IEEE Transactions on Wireless Communications*, 11, 800 –809.

WANG, S., ZHANG, J. et TONG, L. (2012b). A characterization of delay performance of cognitive medium access. *IEEE Transactions on Wireless Communications*, 11, 800–809.

WILLKOMM, D., GROSS, J. et WOLISZ, A. (2005). Reliable link maintenance in cognitive radio systems. *Proc. of 2005 First IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN 2005)*. 371–378.

XIAO, J., HU, R., QIAN, Y., GONG, L. et WANG, B. (2013). Expanding LTE network spectrum with cognitive radios : From concept to implementation. *IEEE Wireless Communications*, 20, 12–19.

YAU, K.-L., KOMISARCZUK, P. et TEAL, P. (2009). Cognitive radio-based wireless sensor networks : Conceptual design and open issues. *IEEE 34th Conference on Local Computer Networks, 2009 (LCN 2009)*. 955 –962.

YIN, S., CHEN, D., ZHANG, Q., LIU, M. et LI, S. (2012). Mining spectrum usage data : A large-scale spectrum measurement study. *IEEE Transactions on Mobile Computing*, 11, 1033–1046.

YOUNIS, O., KANT, L., CHANG, K., YOUNG, K. et GRAFF, C. (2009). Cognitive MANET design for mission-critical networks. *IEEE Communications Magazine*, 47, 64–71.

YUCEK, T. et ARSLAN, H. (2009). A survey of spectrum sensing algorithms for cognitive radio applications. *IEEE Communications Surveys & Tutorials*, 11, 116–130.

ZHANG, Q. et KASSAM, S. (1999). Finite-state markov model for Rayleigh fading channels. *IEEE Transactions on Communications*, 47, 1688 –1692.

ZHAO, Q. et SADLER, B. (2007). A survey of dynamic spectrum access. *IEEE Signal Processing Magazine*, 24, 79 –89.

ZHAO, Q., TONG, L. et SWAMI, A. (2005). Decentralized cognitive mac for dynamic spectrum access. *2005 First IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN 2005)*. 224 –232.

ZHAO, Q., TONG, L., SWAMI, A. et CHEN, Y. (2007). Decentralized cognitive mac for opportunistic spectrum access in ad hoc networks : A pomdp framework. *IEEE Journal on Selected Areas in Communications*, 25, 589 –600.

ZHOU, X., LI, Y., KWON, Y. H. et SOONG, A. (2008). Detection timing and channel selection for periodic spectrum sensing in cognitive radio. *Proc. of IEEE Global Telecommunications Conference, 2008. IEEE GLOBECOM 2008*. 1–5.